May 2023

# Understanding Population Gaps and Using Explainable Machine Learning to Predict Risk of Falls for Senior Adults

Ling Tong
*University of Wisconsin-Milwaukee*

## Recommended Citation

Tong, Ling, "Understanding Population Gaps and Using Explainable Machine Learning to Predict Risk of Falls for Senior Adults" (2023). *Theses and Dissertations*. 3222.
https://dc.uwm.edu/etd/3222

# UNDERSTANDING POPULATION GAPS AND USING EXPLAINABLE MACHINE LEARNING TO PREDICT RISK OF FALLS FOR SENIOR ADULTS

by

Ling Tong

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Biomedical and Health Informatics

at

The University of Wisconsin-Milwaukee

May 2023

# ABSTRACT

## UNDERSTANDING POPULATION GAPS AND USING EXPLAINABLE MACHINE LEARNING TO PREDICT RISK OF FALLS FOR SENIOR ADULTS

by

Ling Tong

The University of Wisconsin-Milwaukee, 2023
Under the Supervision of Professor Jake Luo

Senior adult falls are one of the leading causes of injury and death. The analysis of social determinants and healthcare utilization for senior patients with a history of falls is limited. Most healthcare outcome studies focus on risk factors for falls. There is a lack of studies on patients' socioeconomic and demographic effects on healthcare utilization. With unequal utilization of healthcare adoption, clinical developments, such as clinical decision-making tools, cannot reach people equally, which limits the potential of improving healthcare coverage.

To close the healthcare utilization gap, this dissertation includes two studies focusing on healthcare disparities and the use of innovative technology to provide solutions.

1) To understand the population gaps in healthcare adoption: the first study examines socioeconomic disparities and their impact on healthcare utilization for senior patients with a history of falls in southeast Wisconsin and the Milwaukee metropolitan area, one of the most racial segregated communities in the United States. This study found that disadvantaged social conditions, including under-insurance, residing far from a hospital, lower education, and lower income, were associated with a lower healthcare utilization rate. These findings indicate the need to address healthcare inequities to facilitate equitable care for all patients.

2) To improve the clinical decision support model using an interpretable machine learning framework: The clinical decision support model has been demonstrated to be an effective method to improve healthcare and reduce inequality; however, the mechanisms of the clinical decision support model are difficult to interpret for clinical practitioners. To address this problem, the second study discusses interpretable machine learning uses and applications. The second study discussed an example of machine learning-based text classification for clinical computed tomography reports, specifically for temporal bone fractures, one of the severe outcomes of senior adult falls. This study develops a solution to use an interpretable artificial intelligence framework and computer-based methodology to understand the mechanisms of clinical decision models to close the inequality gap for senior adults. Machine learning models were used to classify fractures based on text reports, and two methodologies were used for interpretation, which resulted in high interpretability, possibly improving the physicians' trust in the clinical decision-making model. This study can assist physicians in using technology more effectively to aid decision-making and increase trust in computerized models. In addition, this study demonstrates machine learning classification models can process a large quantity of clinical texts to reduce the physician's documentation processing workload, possibly leading to high-quality healthcare.

Together, the two papers in this dissertation underscore the importance of addressing healthcare disparities. The study shows it is feasible to use machine learning, a computer-based technology, to form a pathway for better care for patients. By addressing disparities and leveraging technology effectively, it is possible that healthcare providers could provide equitable and high-quality care for all patients, regardless of socioeconomic status or diagnosis.

This thesis is dedicated to my parents,
for the endless love, support, and encouragement

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ACRONYMS

| | |
|---|---|
| EHR | electronic health records |
| ICD | international classification of diseases |
| ADI | area deprivation index |
| RUCC | rural-urban continuum code |
| ZCTA | zip-code tabulation area |
| ACS | american community survey |
| CI | confidence interval |
| OR | odds ratio |
| SD | standard deviation |
| AI | artificial intelligence |
| BOW | bag-of-words |
| NLP | natural language processing |
| CT | computed tomography |
| LIME | local interpretable model-agnostic explanations |
| ML | machine learning |
| SVM | support vector machine |
| TF-IDF | term frequency – inverse document frequency |
| WFS | word frequency score |

# ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support of my major professor, Dr. Zhihui Luo. Dr. Luo’s s mentoring allowed me to succeed in achieving this goal. Dr. Luo guided me through much of my PhD journey and led me to completion. Without Dr. Luo, I could not have gone so far to the end. Dr. Luo has been a constant source of inspiration throughout my PhD journey, and I feel fortunate to have had the opportunity to work with such a brilliant and dedicated researcher. Dr. Luo’s commitment to excellence has set a high bar for the quality of work that I aspire to achieve, not to mention that Dr. Luo challenged me to think deeply and critically about my research. His insightful feedback and constructive criticism have been invaluable in helping me refine my ideas and arguments. I am deeply grateful for the countless projects, papers, and studies that we have worked on together. I believe that Dr. Luo's influence will be lifelong in my future studies, shaping the integrity of my work and inspiring me to continue to strive for excellence.

I also would like to thank my dissertation committee, Dr. Susan McRoy, Dr. Jennifer Fink, and Dr. Min Wu, who graciously shared their insights, advice, wisdom, and precious time to review and critique my work.

I would also like to express my gratitude to my professor when I met in undergraduate school, Dr. Lingyun Luo, for her insightful guidance in steering me towards the field of health informatics. Her recommendation to pursue this path was a pivotal moment in my academic and professional journey, and I will always be grateful for her belief in my potential. Thanks to her encouragement and mentorship, I was able to take a bold leap into this exciting field. It has proven to be the right decision. Her focus on building me strong computer science and programming foundations, which has been invaluable in shaping my work, and I credit her with the solid computer programming skills that have allowed me to excel in health informatics methodological research. I am deeply appreciative of the impact she has had on my education and career.

I would like to extend an additional special thanks to Dr. Timothy Haas for his invaluable guidance and mentorship in my learning of statistical and clinical studies. Without his expert knowledge and patient teaching, I would not have been able to acquire the statistical skills and knowledge that I possess today.

Being a dissertator during the COVID-19 pandemic has been a challenging and unprecedented experience. Not only myself, but all of us have all had to navigate the complexities of balancing our academic pursuits with the demands of a rapidly changing world, all while trying to maintain our mental and emotional well-being. Despite these difficulties, I am proud to have completed a major part of my dissertation during this difficult time.

This work is dedicated to all the people out there who are navigating the healthcare system to get the care they need during this pandemic, as well as to all those who have reached out a helping hand and kept a positive attitude in the face of adversity. Your resilience and strength have been a source of inspiration and motivation for me. I hope that this work will contribute in some small way to the collective efforts to overcome the challenges, the disparity problems during this pandemic, and provide an approach to reduce the inequality in healthcare.

# 1. Introduction

### 1.1 Senior Adult Falls Overview: A Major Public Health Concern

Falls are one of the top causes of injury and death in older adults, which becomes a major public concern.[1], [2] Every year, one in three adults over the age of 65 falls.[3] Over fifty percent of adults over the age of 80 had a history of falls.[2] Twenty to thirty percent of patients with falls suffer from moderate to severe injuries, limiting their ability to live independently in the community. The severe injury of falls requires hospitalization and increases the risk of death[4]. In 2018, there were 2.4 million non-fatal fall injuries among older adults in the United States that required treatment in emergency rooms, with over 700 thousand patients being hospitalized.[5] Over 34 thousand older adults die from unintentional fall injuries. [5] As a result, it is estimated that falls are the fifth leading cause of death in adults over the age of 65,[6] and will become a continuing issue in senior adult care.

Because of the high prevalence of age-related physiological changes, comorbidities, and delayed functional recovery, older people are more vulnerable to injury, leading to further deconditioning and more falls [1]. While 30 to 50 percent of falls result in minor injuries, such as bruises or lacerations, five to ten percent of falls result in major injuries, including fractures and traumatic brain injuries [7]. Falls are the leading cause of traumatic brain injury in older people, accounting for 46 percent of fall-related deaths in the cohort [8]. Seventy-six percent will slightly lose mobility, and 50 percent will lose the ability to perform activities of daily living. Falls are also the leading cause of injury-related death in older adults [9]. The risk of death from fall-related injuries is 2.33 times higher than the risk of death from other causes [10]. Approximately 50 percent of older adults who fall are unable to stand up and remain on the ground [11]. Pressure sores, pneumonia, and dehydration can result from the inability to stand up. Approximately one-half of older adults will fall again within six months [12]. This enables many older adults to have a fear of falling [13]. Research indicates that up to 40% of them will limit their daily activities to prevent further falls [14]. Restricted activity can lead to social isolation and depression. Further declines in physical fitness and social isolation, in turn, increased the risk of falls [15].

As the United States population is projected to reach 80.8 million by 2040 and 94.7 million by 2060 [16], senior adults are one of the fastest-growing demographics. The prevalence of falls and fall-related injuries in this demographic is projected to escalate, placing substantial strain on healthcare systems. Therefore, it is essential to provide an overview of the elderly fall problem to improve healthcare.

The etiology of falls in older adults is multifactorial, encompassing intrinsic factors, such as age-related declines in muscle strength, balance, and cognitive functioning, as well as extrinsic factors, such as environmental hazards and medication side effects [7]. Furthermore, chronic conditions, such as osteoporosis, arthritis, and Parkinson's disease, have been identified as contributing factors to the increased risk of falls in the elderly population [17]. In light of these complex interactions, it is imperative to develop a comprehensive understanding of the underlying causes and potential interventions to effectively address this growing public health concern [18].

Despite the considerable body of research on falls in older adults, there remain significant gaps in knowledge related to the most effective prevention strategies and interventions [19], [20]. While certain measures, such as strength and balance training, home safety modifications, and medication reviews, have demonstrated efficacy in reducing fall risk, the optimal combination of interventions and their applicability to diverse populations are not yet fully understood [21]. Additionally, the implementation of such interventions in real-world settings is often hindered by barriers such as inadequate funding, a lack of professional training, and insufficient collaboration between healthcare providers and community organizations [22]. The economic burden of falls among older adults is substantial, with direct medical costs reaching billions of dollars annually [23]. This financial strain is compounded by indirect costs, including lost productivity, informal caregiving, and reduced quality of life for both patients and their caregivers [24]. As a result, investing in research and policy initiatives focused on fall prevention is not only a moral imperative but also an economically prudent strategy to mitigate the long-term consequences of this public health challenge [25].

In conclusion, the issue of falls among older adults is a complex and multifaceted problem that warrants a comprehensive and interdisciplinary approach. By addressing the existing gaps in knowledge and overcoming barriers to the implementation of evidence-based interventions, it is possible to make significant strides in reducing the incidence and impact of falls in this rapidly growing population. Ultimately, this will contribute to the promotion of healthy aging, the preservation of independence and quality of life, and the optimization of healthcare resources. Upon investigation, there are major gaps in the treatment, diagnosis, and prevention of falls in clinical care. Identifying these gaps can significantly help find a solution to address them and optimize health care resources. Providing a clear background investigation also provides a clear path for new healthcare improvements for elderly fall prevention to reduce potential elderly falls [20].

### 1.2 Inequality of Socioeconomic Conditions and Healthcare Utilizations

The risk factor of falls has been identified from a clinical perspective, with many comprehensive evaluations available [9], [20], and [26]. The major two factors are related to person-specific factors and environmental factors. Environmental factors include poor-fitting footwear, slippery floors, tripping hazards, loose rugs, a lack of stair railings, poor lighting, and so on [27]. Improving the environment has been demonstrated as an effective method to reduce fall risk [28]. Public health studies have come up with interventions to measure in-home fall risk detection, walking path planning, and robot-based alerts and preventions. Personal factors include individual characteristics such as age, functional abilities, chronic diseases, and gait disturbances. [1]. Most risk factors found in a literature review were intrinsic and included a wide range of risk categories: demographic profile, lower extremity strength, vertigo and dizziness, vision, cognition, cardiovascular disease, medications, depression, gait, and balance caused by normal aging and pathological effects. Additionally, each risk category had several risk factors that, when co-existing, might increase the chances of falling. [29], [30] For instance, orthostatic hypotension, hypertension, and atrial fibrillation fall under the cardiovascular disease category. For example, Gangavati et al. found that older adults with systolic orthostatic hypotension and uncontrolled hypertension had a higher risk of falling than those with uncontrolled hypertension alone. [31]

While fall risk factors have been comprehensively studied, socioeconomic factors have received less attention. [32] Those with lower socioeconomic status may not have equal access to care following a fall. [33] The socioeconomic status of the elderly refers to an individual's social standing, which is typically measured by several indicators such as education, occupation, income, and location. [32] Because higher socioeconomic status is associated with better well-being and lower nutritional risk, older people with lower socioeconomic status may be more vulnerable to falls. [34]. Despite this, very few studies have looked into the relationship between the elderly's fall risk and socioeconomic factors, and it is unknown how socioeconomic factors affect the patient’s healthcare adoption rate for patients with falls.

Health equity is a significant gap to ensure all patients can receive equitable healthcare. Offering equitable health care leads to more efficient healthcare systems overall, as a healthier population requires less medical care. [35] For patients with lower socioeconomic status, there might be a chance they received a lower level of healthcare compared to the average. [36] Investigating the gaps in health disparities in the population can benefit a large population, especially those of lower socioeconomic status. [37] This study aims to discover the equity problems in healthcare utilization and provide guidelines, ensuring different outcomes and recommendations for specific populations. Guidelines can potentially help practitioners address inequitable variations in health care and the allocation of services, which might result in health disparities [36, 37].

Falls among older adults, while already a significant public health concern, are further compounded by disparities in socioeconomic status and access to healthcare [38]. Although numerous studies have focused on the clinical and environmental aspects of falls, there remains a paucity of research examining the interplay between socioeconomic factors and the risk, prevention, and management of falls in older adults [39]. This gap in knowledge hinders the development of equitable interventions and policies, ultimately exacerbating health disparities and perpetuating cycles of disadvantage among vulnerable populations [35].

Socioeconomic factors, such as education, income, and neighborhood characteristics, have been shown to influence health outcomes and access to healthcare services in older adults [40]. For instance, those with lower levels of education may have limited health literacy, resulting in suboptimal engagement with healthcare providers and reduced adherence to fall prevention strategies [41]. Similarly, older adults residing in socioeconomically disadvantaged neighborhoods may face additional barriers to accessing appropriate healthcare services, including transportation difficulties, inadequate health insurance coverage, and a lack of culturally competent providers.

Moreover, socioeconomic disparities may influence the availability and quality of resources within the home and community environments, further exacerbating the risk of falls [42].

Older adults with limited financial means may be unable to afford necessary home modifications or assistive devices, such as grab bars and walkers, which have been demonstrated to reduce fall risk [28]. Additionally, those residing in underprivileged communities may experience higher levels of environmental hazards, such as poorly maintained sidewalks and inadequate public lighting, further increasing the likelihood of falls and associated injuries [43].

Addressing the issue of health equity in the context of falls among older adults requires a multifaceted approach that considers the interplay between clinical, environmental, and socioeconomic factors [44]. This may involve the development and implementation of targeted interventions that address the unique needs and challenges faced by socioeconomically disadvantaged older adults, as well as broader policy initiatives aimed at reducing health disparities and promoting equitable access to healthcare services [39]. Such strategies may include the provision of affordable and accessible home modification programs, the expansion of community-based fall prevention initiatives, and the promotion of cultural competence among healthcare providers [45]. Furthermore, the integration of social determinants of health into fall risk assessments and care plans can facilitate the identification of at-risk individuals and the tailoring of interventions to their specific needs.

In conclusion, the examination of socioeconomic factors in relation to falls among older adults is a critical yet underexplored area of research with significant implications for health equity and the optimization of healthcare resources. By addressing these gaps in knowledge and implementing targeted interventions and policies, it is possible to promote more equitable healthcare outcomes and foster the well-being of all older adults, regardless of their socioeconomic status.

### 1.3 Senior Adult Falls and Bone Fractures

Studies show older adults' falls are associated with bone fractures, a severe outcome with one of the worst prognoses. [46] Approximately 30% of falls result in an injury that requires medical attention, and fractures occur in approximately 10% of total falls. [47] People over the age of 65 experience significantly increased morbidity and mortality [2]. In addition, falls cause 90 percent of all hip fractures, significantly increasing the death rate of older adults [48]. In the first year after a hip fracture, 20 percent of older patients will die. The presence of a bone fracture has been linked to 40% of deep vein thromboembolism and 10% of delirium. [49] The complications resulted in longer hospital stays, increased mortality, and the risk of nursing home placement. The death is primarily due to pneumonia, cardiac disease, pulmonary embolism, and surgical complications. [18] It is estimated that women before the age of sixty tend to extend their arms as they fall, which increases the risk of forearm fractures. Women tend to fall sideways after that age and have a higher incidence of hip fractures [14].

Falls and their consequences account for a significant portion of avoidable health care costs. In the United States, the cost of medical resources spent on falls and related injuries is estimated to be close to $30 billion in 2018 [23]. In the United Kingdom, the cost of falls is 1.6 billion pounds [50]. As the global population ages, expenditures are expected to approach $55 billion by 2030. [51] Furthermore, fall-related medical events account for 40% of nursing home placements, contributing to even higher healthcare costs. [52] The national costs of fall-related treatments account for 0.85 to 1.5 percent of total healthcare expenditures [53].

An issue with bone fracture treatment is that it requires significant medical resources [54]. Fractures frequently necessitate multidisciplinary care, including visits to the emergency department, radiology, and orthopedics, which consume significant medical resources on documentation and take a large amount of time for clinicians [54]. The documentation increases the burden on the health care system. In this situation, it has been demonstrated that the burden of clinical documentation on professionals is the major cause of negative effects on health care. Reducing the documentation burden on U.S. clinicians is an urgent priority within the healthcare community [55]. For example, the American Medical Informatics Association, the leading association for health care administration, has prioritized the documentation overload problem [566]. **The** American Medical Informatics Association has taken leadership in the 25x5 initiative to address this nation-wide documentation issue. This initiative leverages the collective expertise of key stakeholders in health care, industry, and policy to prioritize and implement the mission of the initiative. The mission is to reduce the documentation burden for U.S. clinicians to 25% of the **current level** in the next 5 years.

### 1.4 Using Explainable Machine Learning to Reduce the Burden of Documentation

Elderly falls are controllable. [57] There are predictive patterns for them based on known risk factors and defined demographics. A study has shown that the physiological changes associated with aging account for 72% of falls, whereas the other 18% are unpredictable and categorized as accidental because they are the result of environmental hazards [1]. The fact that cognitive and motor performance deficiencies are significantly associated with fall risk suggests that falls can be predicted through clinical assessments. Conventionally, the most accurate health monitoring occurs in a laboratory or hospital setting, but such hospital-based health monitoring is prohibitively expensive and not regularly undertaken. [58] The automatic fall detection approach, on the other hand, allows for the early detection of elderly individuals in danger of falling or the detection of those who have already fallen so that subsequent interventions can be made. This capability can reduce the incidence of initial and subsequent falls and mitigate physical and mental suffering. Therefore, machine learning-based methods utilizing known risk factors and demographics can be developed to predict the risk of falls, and a fall prevention program based on the prediction model has the potential for clinical intervention.

Currently, several machine learning detection and risk assessment models have been developed for the prediction of future falls in individuals without a history of falls. [59]–[64] The studies demonstrate that machine learning has the potential to be applied to a wide variety of medical applications, including but not limited to clinical decision-making. A large number of successful artificial intelligence applications in healthcare have been used in prediction tasks, such as predicting adverse events [65], [66], drug responses [67], complications [68], and medical diagnoses [69], [70]. While many fall detection algorithms have been developed, the prediction accuracy must still be approved to fully realize the model's potential in clinical practice. There is a foreseeable future when artificial intelligence can fully leverage electronic health records, which can develop complex models to assess the risk of a patient based on their medical history [71]. Therefore, further studies of AI in fall prediction can provide clinicians with alerts for high-risk patients, reducing the occurrence of future falls [72].

Electronic health records have been acknowledged as a key to improving healthcare quality [73]. Computerized decision-making models are widely used in clinical applications for disease discovery, identification, and prediction [74]. However, most current studies use structured features to build models. Unstructured data, such as free-text clinical notes, is rarely used. The limited use of free-text data is due to format issues [75], [76]. For example, clinical texts require human-level intelligence to process complex linguistic rules that go beyond the scope of simple classification. To leverage clinical texts and build an accurate model, a common method is to label the clinical text. The manual process of creating free-text clinical notes and labels was inevitably expensive. The cost limits the wider use of free-text clinical notes.

Natural language processing [76] techniques are commonly used to build clinical classification models using free texts. Natural language processing mimics how humans learn a language by comprehending its semantics. Understanding natural language requires linguistic knowledge such as morphology, syntax, and pragmatics [77]. We have seen considerable progress in natural language processing and AI-based clinical decision-making classifiers [73]. However, understanding a model's mechanism requires extensive computer-domain knowledge [78]. Clinical practitioners need a simple method to understand the mechanisms of decision-making models.

Even though machine-learning clinical classification models have improved, only a few of them are used in clinical settings because doctors don't trust them [79]. A validation set is a common way to ensure that a machine learning classification is generalizable [80]. However, a validation set must not replace real clinical contexts. Before using a computerized model in clinical practice, physicians must be confident that the decision-making model is applicable to patients in clinical settings. It is impossible to establish trust unless physicians understand how a model makes decisions based on medical domain knowledge. The lack of trust and transparency in decision-making models raises concerns about making incorrect decisions [75].

It is possible to address this distrust by visualizing classifier interpretations. [81] A set of untemplated narrative reports from temporal bone computed tomography was used in our case study. These reports differentiate between those with and without fractures. Our visualization demonstrates that many aspects of clinical texts, including word frequency and word selection, impact the final classification decision. For example, we demonstrate that the presence of some words, such as ‘fracture’, is the reason a classifier makes a positive classification. Using our visualization, physicians can combine medical domain knowledge with visualization to assess the validity of the highlighted keywords. Therefore, we believe that our visualization can boost physicians’ confidence in using classification models. This study could accelerate the adoption of machine learning-based decision-making systems in clinical settings and reduce the documentation burden for clinical practitioners.

### 1.5 Problem Statement

This dissertation includes two studies to address the two gaps in the topic of older adults’ falls. The first study starts with association studies to identify the socioeconomic gaps influencing health care adoption in the southeastern Wisconsin area. The second study's goal is to reduce clinicians’ documentation burden and build trust using interpretable machine learning models. The second study developed a clinical document documentation classification system to process natural language text and make an automatic classification of temporal bone fractures. The evaluation shows the interpretable framework can successfully highlight the most impactful word that contributes to the classification results. Therefore, the machine learning model has the potential to be deployed in a hospital environment to process clinical documents. The explainable model can also increase clinicians’ trust in clinical decision-making tools. A clinician’s trust may form a pathway to higher healthcare quality using the automated tool.

#### 1.5.1 Investigating the Inequality of Health Care for the Older Adult Population

Despite many studies on fall risk factors, physicians at Froedtert Hospital reported that hospital admissions due to falls appear not to be proportional to race and gender distribution. Therefore, the patient demographic may be a factor in fall-related care. Also, patients with low socioeconomic status cannot receive the same level of care as those with high socioeconomic status, which may exacerbate health care disparities. To uncover the inequality issue for patients with fall diagnoses, an association study investigating socioeconomic variables that involves fall-admitted students is required. These socioeconomic factors, such as the patient's location, income, and type of insurance used, will play an important role in health care adoption following a patient's fall. The study can help to understand how socioeconomic factors influence patients with a history of falling, revealing potential disparities in health care access. It is also important to understand the relationship between socioeconomic factors and fall risk factors, which can help identify the most vulnerable population and provide targeted interventions to reduce the risk of falls. Understanding the causes of disparity and promoting a solution to improve equal access to care for all patients is critical.

### 1.5.2 A Transparent Text Classifier Can Reduce the Documentation Burden

The burden of clinical documentation on professionals has had a proven negative impact on health care [82]. This burden leads to a variety of negative outcomes, including clinician burnout and decreased job satisfaction, increased medical errors, and hospital-acquired conditions. The use of computed tomography (CT) documentation to diagnose fall-related fractures has resulted in a large number of documents and highly specialized narratives. To reduce the burden, machine learning (ML) can successfully leverage free-text clinical notes to classify patients' diagnostic outcomes. However, the interpretation of a classification result remains challenging due to the model’s complexity. A complex machine learning model is often called a black box, which causes difficulties in understanding the mechanism. To increase the transparency of the machine learning model even further, we used word frequency analysis and a comprehensible text explainer to provide simple visualizations of how the machine learning model makes predictions and ultimately to improve the mutual trust of computerized models, finally leading to the adoption of automatic computer-based systems in clinical practice.

### 1.6 The Contribution of Two Studies

The first contribution is on the findings of healthcare disparities. We investigate socioeconomic disparities and their impact on healthcare utilization in Southeast Wisconsin for senior patients with a history of falls. We designed a cohort study and acquired data from the electronic health records of Froedtert Health Care Center. We analyzed the association between elderly falls and medications, vital signs, a list of diagnoses, and a number of socioeconomic factors, including age, education, race, insurance coverage, and individual income. This study discovered that disadvantaged social conditions, including a lack of insurance, living far from a hospital, a lack of education, and a lower income, were associated with lower healthcare utilization. These findings highlight the importance of addressing healthcare inequities in order to provide equitable care to all patients. The statistical methodology of the study can also be applied to discover the association between socioeconomic status and healthcare utilization problems in other specialties and diagnoses. Using this method, we have extended our studies into other aspects of healthcare, including the inequality of telemedicine visits [83], [84], obesity care [85], tertiary rhinology care [86], idiopathic sudden sensorineural hearing loss [87], treatment of dysphonia [88], Meniere’s disease [89], unilateral vocal fold paralysis [90], vestibular schwannoma [91], and post-tympanotomy tube otorrhea [92]. Therefore, we have demonstrated the methodology's broad applicability in finding healthcare inequality from electronic health records.

The second contribution emphasizes the significance of creating interpretable machine learning models for clinical text classification, specifically in the context of clinical computed tomography (CT) reports concerning temporal bone fractures. By developing an automated text classification system to classify patients' diagnostic outcomes for bone fractures, this study seeks to reduce the spending of healthcare resources on physician documentation, which can lead to improved efficiency in healthcare delivery. Utilizing a large dataset of CT reports, the machine learning model is trained to classify patients' diagnostic outcomes accurately. The incorporation of word frequency analysis and a comprehensible text explainer provides simple visualizations of the model's decision-making process. This increased transparency fosters better understanding and trust in computerized models among physicians and other healthcare professionals.

The significance of the second research lies in its potential to help physicians use technology more effectively to aid decision-making and reduce documentation workload, allowing them to focus more on patient care. By increasing trust in computerized models through interpretable machine learning, healthcare professionals can become more receptive to the adoption of advanced technology in clinical practice. Ultimately, this can lead to more accurate and efficient diagnoses, improved patient outcomes, and optimized healthcare resources.

# 2. Related Work

## 2.1 Senior Adult Falls Problem and Clinical Outcomes

Falls are a major clinical problem among older adults, contributing to significant morbidity and mortality. In fact, falls are the leading cause of injury-related death and hospitalization among adults aged 65 years and older [93]. Falls can lead to serious injuries, such as hip fractures, head injuries, and lacerations, and can result in long-term disability or even death. Falls can also have a psychological impact on older adults, leading to fear of falling, social isolation, and decreased quality of life [94]. Given the significant impact of falls among older adults, researchers have focused on understanding the factors that contribute to falls and developing effective prevention strategies. Several risk factors have been identified for falls among older adults, including advanced age, female gender, cognitive impairment, visual impairment, and mobility impairment. [95], [96] Additionally, social determinants of health, including socioeconomic status, race, and geography, may also contribute to falls among older adults [97].

To address the clinical problem of senior adult falls, healthcare providers have implemented several fall prevention strategies. These strategies may include exercise programs, medication management, environmental modifications, and patient education [21]. However, the effectiveness of these interventions can be limited by a lack of healthcare access or other social determinants of health. Machine learning has emerged as a potentially powerful technique for identifying risk factors for falls among older adults and developing personalized fall prevention strategies. Researchers have used machine learning to predict falls among older adults [98], identify factors that contribute to falls [99], and develop personalized rehabilitation plans for fall victims [100]. These studies suggest that machine learning may be a promising approach for addressing the clinical problem of senior adult falls.

## 2.2 The Socioeconomic Inequality Gaps in Healthcare Adoption

In addition to being a major clinical problem, falls among older adults also highlight the issue of healthcare inequality. Socioeconomic factors, including income, education level, and insurance coverage, can significantly impact healthcare adoption and utilization among older adults [101], [102]. Older adults with limited financial resources or inadequate insurance coverage may face barriers to accessing necessary healthcare services, such as assessment and management of fall risk.

Previous studies have demonstrated that socioeconomic disadvantage is associated with increased fall risk among older adults [103], [104]. Additionally, studies have shown that socioeconomic factors also impact healthcare utilization among older adults with falls. For example, older adults with lower income, lower education levels, and inadequate insurance coverage were less likely to access rehabilitation services and were more likely to experience poor outcomes after a fall [105]. These studies suggest that addressing socioeconomic factors is crucial for reducing healthcare inequality in the management of senior adult falls.

To address the issue of healthcare inequality in the management of senior adult falls, healthcare providers and policymakers have implemented several strategies. For example, healthcare providers can use telemedicine to improve access to healthcare services for older adults with limited mobility or inadequate insurance coverage [106]. Additionally, community-based interventions, such as fall prevention programs, transportation services, and healthcare navigation services, may also help reduce healthcare inequality and improve healthcare utilization among older adults with falls. [107], [108] Machine learning has also shown promise for addressing healthcare inequality in the management of senior adult falls. Machine learning algorithms may help identify socioeconomic factors associated with fall risk, improve identification of high-risk individuals, and develop targeted interventions for those at highest risk [109]. By integrating socioeconomic factors into machine learning algorithms for fall prevention, healthcare providers can help address healthcare inequality in the management of senior adult falls.

## 2.3 Reducing Physicians’ Workload May Close the Inequality Gap

One potential approach to reducing healthcare inequality and improving healthcare access and quality is to leverage technology to reduce physicians' workload. Many healthcare providers, particularly in underserved areas, face a high patient-to-provider ratio, which can lead to burnout and reduced quality of care. Furthermore, the COVID-19 pandemic has emphasized the importance of telehealth and digital health technologies in addressing healthcare access issues, particularly for vulnerable populations. [110]

Several studies have found that telehealth and digital health technologies can reduce physician workload and improve healthcare access and quality. For example, telehealth consultations have been found to be equivalent to in-person consultations in terms of diagnostic accuracy and patient satisfaction [111]. Digital health technologies, such as mobile health apps and patient portals, can also improve healthcare access and quality, particularly for patients with chronic conditions [112].

Moreover, studies have shown that reducing physician workload through the use of technology can help address healthcare inequality by improving access to care for underserved populations. For example, a study by Reed [113] found that telehealth consultations improved access to specialist care for patients in rural areas. Another study found that a telehealth program for diabetic patients led to improved glycemic control among low-income and Hispanic patients [114].

Machine learning can also be leveraged to reduce physician workload and improve healthcare access and quality. For example, machine learning algorithms have been used to identify patients at high risk for adverse health outcomes, such as hospital readmissions, and develop personalized care plans [115]. Additionally, natural language processing (NLP) techniques can be used to extract valuable information from unstructured clinical notes, reducing the time and effort required for physician documentation.

Overall, reducing physician workload using technology, including telehealth, digital health technologies, and machine learning, can help address healthcare inequality and improve healthcare access and quality. By increasing efficiency and reducing barriers to care, these approaches can help ensure that all patients receive the high-quality care they need and deserve.

### 2.4 Clinical Decision-Making Tools Can Reduce the Burden and Improve Care Quality

Managing clinical documentation is a critical aspect of healthcare delivery, but it can also be a significant burden on providers, leading to burnout and reduced quality of care. Many healthcare systems struggle with clinical documentation overload, with providers feeling overwhelmed by the amount of documentation required for each patient encounter. This problem can be particularly acute in resource-limited settings, where there are fewer providers to handle a larger caseload [116].

Machine learning has the potential to alleviate some of the burden of clinical documentation and improve healthcare quality by automating parts of the documentation process. For example, machine learning algorithms can be used to automatically extract information from electronic health records (EHRs) and other clinical documents, reducing the time and effort required for providers to manually input this information [117]. Additionally, machine learning algorithms can be used to identify high-risk patients, predict clinical outcomes, and develop personalized treatment plans, all of which can enhance the quality of care delivered to patients [118].

One promising application of machine learning in clinical documentation is the development of automatic decision-making tools that can help providers make more accurate and informed decisions in real-time. These tools can be integrated into EHRs or other clinical systems and can provide real-time recommendations based on patient data and other relevant clinical information. For example, a machine learning algorithm could be used to flag patients who are at high risk for medication errors or adverse drug reactions, alerting providers to potential issues before they occur [119]. Similarly, machine learning algorithms can be used to classify clinical notes and extract relevant information, allowing providers to access key information more efficiently [120].

Overall, using machine learning to develop automatic decision-making tools can help reduce the clinical documentation load and improve healthcare quality. By automating parts of the documentation process and providing real-time decision support, providers can focus on providing high-quality, patient-centered care rather than struggling with the administrative burden of clinical documentation. These tools can also reduce errors and improve clinical outcomes, ultimately benefiting patients and providers alike.

### 2.5 Clinical Text Classification

To reduce error and improve efficiency, many studies have begun to explore the adoption of text classification systems since the 1990s. [121]. Early studies focused on rule-based methods to build classifiers for medical documents [121]. For example, Aronow et al. [122] developed

NegExpander, a computerized system that distinguishes between positive and negative evidence in radiological reports. The system recognizes nouns and conjunctive phrases that define negation boundaries. The proposed classifier had a precision value of 93%. Thomas et al. [123] developed a text search algorithm based on association rules and implemented a computerized text classification system. The fully computerized way that radiographic reports were categorized as "normal," "neither normal nor fracture," and "fracture". A rule-based system is a simple and effective AI-based application. However, the speed and ability to handle complex tasks are limited.

On the other hand, ML-based classifiers can adjust their parameters to adapt to the ever-changing word usage in medical documents. In recent years, machine learning studies have begun to use complex statistical models to classify clinical texts. In decision-making, Bayesian networks [124]–[129], support vector machines [127], [130]–[133], and decision trees [125], [128], [134]–[137] have been widely used. These models outperformed the rule-based system in terms of classification accuracy. In 2006, de Bruijn et al. [127] used supervised machine learning approaches to develop classifiers that automatically detect acute wrist fractures in radiological reports. They reported that the support vector machine (SVM)-based text classifier performed best overall, with 94% accuracy. Zuccon et al. [132] experimented with feature engineering in SNOMED CT concepts to improve medical image classification accuracy. The classifier developed by Guido Zuccon et al. could correctly identify fractures from radiological reports. It is also stated that when using bigram or SNOMED+bigram features, the Nave Bayes classifier had the highest F1-score. Dai et al. [138] created a classifier for bone fracture detection using regular text classification in 2017. Topic modeling and document similarity measurement are used to train the classifier.

The lack of transparency in the classifier remains an unresolved issue. As a result, physicians struggle to understand why the classifier makes positive or negative classifications. A recent study [139] used a large dataset to implement name entity recognition and bone fracture classification. There are some attempts at machine learning models' interpretation of clinical texts. For example, a recent study [126] built five machine learning algorithms to classify Alzheimer's drugs' mechanisms of action. The author visualized a decision tree and tried to provide some textual interpretations. Obviously, more attempts are needed to fully reveal how machine learning models interpret the classification results. In these studies, the model's interpretability issues have not been fully resolved. To help people understand how decision-making systems work, it is important to build an interpretable model that is clear and easy to understand.

## 2.6 Interpretable Machine Learning

Methods based on machine learning are effective for classifying free text reports. An ML model, as opposed to a rule-based system, consists of an algorithm that can learn latent patterns without hard-coding fixed rules [140]f an algorithm that can learn latent patterns without hard-coding fixed rules [140]. One disadvantage of machine learning models is the difficulty of interpreting classification results [141]. To address this weakness, recent studies have begun to interpret machine learning models. This field of study is known as "interpretable machine learning" [141]. An ideal solution for interpretable machine learning is to provide the evidence and reasoning for the user. Furthermore, users can discover knowledge and justify predictions based on provided evidence [142]. Therefore, interpretable machine learning models increase user trust in classifiers. Researchers have developed two types of model interpretation techniques: model-agnostic and model-specific approaches [143]. The model-agnostic approach explains the prediction of an ML model by approximating the output of the model's algorithms. Shapley Values, Independent

Conditional Expectation Plots, Local Interpretable Model-Agnostic Explanations, Permutation Feature Importance, and Partial Dependence Plots are a few examples [143]. Model-specific explanation methods, on the other hand, excel at explaining complex models like tree ensemble models and artificial neural networks [142]. There is also open-source software available, such as SHAP [144], Eli5 [145], and InterpretML [146]. These tools can perform a variety of tasks, including image and text classification. Interpretable machine learning has recently been used in clinical practice for a variety of medical applications, such as predicting mortality risk [147], [148], predicting abnormal ECGs [149], and finding different symptoms from radiology reports that suggest limb fracture and wrist fracture [127], [132]. These studies have demonstrated the potential of interpretable machine learning in medical applications.

Interpretable machine learning models may be able to provide accurate predictions while also being interpretable, which can increase user trust and improve the understanding of the underlying features that influence the predictions. One study used interpretable machine learning to predict the risk of developing heart disease [149]. The authors developed a model that used patient data to predict the risk of developing heart disease in the next ten years. The model was made interpretable by using SHAP values to identify the features that contributed the most to the prediction. The study found that interpretable machine learning models can be used to develop accurate and interpretable models for predicting heart disease.

Another study used interpretable machine learning to identify patterns in electronic health records (EHRs) to predict the risk of hospital readmission [150]. The authors developed a model that used patient data from EHRs to predict the risk of readmission within 30 days. The model was made interpretable by using SHAP values to identify the features that contributed the most to the prediction. The study found that interpretable machine learning models can be used to develop accurate and interpretable models for predicting readmission risk. In addition, one study used interpretable machine learning to classify diabetic retinopathy from fundus photographs [151]. The authors developed a model that used machine learning to classify retinal images as positive or negative for diabetic retinopathy. The model was made interpretable by using a combination of gradient-based and model-based methods to identify the features that contributed the most to the classification. The study found that interpretable machine learning models can be used to develop accurate and interpretable models for classifying diabetic retinopathy.

Despite the value of model interpretability in machine learning models, model interpretation developments are still in the preliminary stages. There is a significant gap between physicians’ desire to understand the prediction and the model’s lack of interpretability. In response to this need, we developed a study investigating interpretable classification models in radiological texts the preliminary stages. There is a significant gap between physicians' desire to understand the prediction and the model's lack of interpretability. In response to this need, we developed a study investigating interpretable classification models in radiological texts. The research was divided into two parts: First, we created a text classifier to classify text radiological reports automatically. Then we conducted model interpretations at the text level. We explored how keywords affect model classification results. To the best of our knowledge, this is the first study that interprets classification results based on temporal bone CT reports.

The study found that interpretable machine learning can be used to create effective text classifiers in radiological reports. The model-agnostic approach was used to interpret the text classifier, with the partial dependency plot and permutation feature importance methods being particularly effective in identifying key features in the data. The study also found that the use of interpretable machine learning models increased user trust in the classification results.

Furthermore, we found that certain keywords were more strongly associated with certain classifications. For example, the presence of the keyword "fracture" was strongly associated with a classification of "positive for fracture". This information can be used to improve the text classifier by including these keywords as features in the model.

Overall, we demonstrate the potential of interpretable machine learning in the field of radiology and suggest that further research in this area could lead to improved text classification models and increased understanding of the underlying features that influence classification results.

# 3. Study One: Socioeconomic Gaps of Senior Adult Falls and Utilizations

Abstract: To examine the social determinant factors of healthcare utilization for senior patients with a history of falls We analyzed the effects of socioeconomic factors on the utilization rate of healthcare in a tertiary care center, including 495,041 senior adults. We included zip code tabulation areas to measure socioeconomic factors on a community level. Cohort group comparison and multiple linear regression models evaluated the association between healthcare service utilization and age, sex, education, race, insurance type, distance, and income levels. The result shows patients with a history of falls were older than those without a history of falls (79.4, standard error = 12.1 vs. 75.4, standard error = 11.6 years old), predominantly female (odds ratio [OR]: 1.26, 95% confidence interval [95% CI]: 1.24-1.28), and white (OR: 1.35, 95% CI: 1.32-1.38). Patients with a fall history were predominantly retired (OR: 2.53, 95% CI: 2.49–2.58), publicly insured (OR: 2.88, 95% CI: 2.82-2.93), and more likely to require an interpreter during a hospital visit (OR: 2.40, 95% CI: 2.35-2.44). Using a geospatial analysis, healthcare utilization was higher in areas close to the care center. A regression model showed that community median income, private insurance rate, and college education rate were positively associated with healthcare utilization. We conclude that lower utilization of healthcare is associated with disadvantaged neighborhood social conditions, including under-insured status, residing far from a hospital, lower education, and lower income. We revealed the current inequities and disparities in the healthcare of senior adult fall patients in Southeast Wisconsin.

## 3.1 Methodology

The objectives of this study are as follows: first, to investigate the relationship between socioeconomic status and fall risk in the elderly population. Second, to examine the healthcare utilization rate and healthcare disparities among elderly fall patients with different socioeconomic statuses. Third, to identify the gaps and inequities in healthcare utilization for elderly fall patients and provide recommendations for improving healthcare delivery and equity. Finally, to develop guidelines for practitioners to address inequitable variations in healthcare and allocate services to reduce health disparities among the elderly fall population,

To understand the health equity challenges and close the gap for senior adult patients, this study aimed to evaluate the impact of socioeconomic factors upon access to healthcare facilities for patients older than 60 years old, with and without a history of falls. This study employed association analysis to determine how health care utilization is related to a variety of socioeconomic factors. The study was conducted in a health system within the Southeastern Wisconsin area. The investigated facility currently serves the majority of the elderly population residing in Southeast Wisconsin, in the United States of America. This study would be helpful to provide recommendations for healthcare policymakers and providers in order to address the inequities and disparities in healthcare for senior adult fall patients.

This study population included all patients aged 60 or older who received care at Froedtert Hospital and Medical College of Wisconsin between March 1, 2020, and March 1, 2022. The electronic health record is acquired through the Clinical Translational Science Institute. The Clinical Research Data Warehouse maintains a database of the Froedtert and Medical College Electronic Health Records. The database currently contains 2.3 million individual patient records. Clinical data is available upon registration for institutional members at the Clinical and

Translational Science Institute. Non-institutional researchers may register as community members for data access. The study was approved by the Institutional Review Board of the Medical College of Wisconsin.

This study identified patients with a history of falls using diagnostic codes in their electronic health records and classified patients into two groups: fall patients and non-fall patients. Fall patients were defined as those who had at least one diagnosis code for a fall-related injury during the study period. Non-fall patients were defined as those not diagnosed with falls and related injuries during the study period.

This investigation collected demographic and clinical data, including age, gender, race, insurance type, co-morbidities, and medication use, for all patients. To evaluate socioeconomic status, this study also collected socioeconomic data at the zip-code tabulation area (ZCTA) level, including median household income, educational attainment, and rural and urban residence. Multivariable logistic regression models were employed to examine the association between socioeconomic factors and healthcare utilization for fall patients compared to non-fall patients, together with determining odds ratios (ORs) and 95% confidence intervals (CIs) to quantify the association. The zip-code area-based variables were from the United States Census Bureau’s 2018 American Community Survey. The data linking to the zip-code tabulation area allowed a community-level socioeconomic analysis throughout the southeastern Wisconsin area. The multiple regression model was adjusted by patients based on age, gender, race, insurance type, and co-morbidities to examine potential effects.

### 3.1.1 Study Cohort and Patient-Based Variables

The study cohort included all senior adult patients (>60 years old) who have visited the hospital and registered as in-person visits between 2020 March and 2022 March. This study split the patients into two exclusive groups. Patients with a history of falls, defined as having at least one fall-related ICD-10 diagnosis in Table 1, during the study period, were identified as the falls group. Patients with non-fall-related diagnoses in their electronic health records were classified into the non-fall group. Within each group, this study collected the age, gender, race, insurance, ethnicity, employment status, and interpreter assistance records during a visit. Age was calculated as the date of visit minus the date of birth for each patient. Sex, race, ethnicity, employment status, and the requirements of an interpreter during a clinical visit were acquired from electronic health records. Insurance status was classified into public, private, other, and uninsured based on the payer's information from the database.

Table 1. List of ICD-10 codes for the Falls Group

| ICD Code | ICD Type | Description |
|---|---|---|
| W00 | ICD-10 | Fall due to ice and snow |
| W01 | ICD-10 | Fall on same level from slipping, tripping and stumbling |
| W03 | ICD-10 | Other fall on same level due to collision with another person |
| W04 | ICD-10 | Fall while being carried or supported by other persons |
| W05 | ICD-10 | Fall from non-moving wheelchair, nonmotorized scooter and motorized mobility scooter |
| W06 | ICD-10 | Fall from bed |
| W07 | ICD-10 | Fall from chair |
| W08 | ICD-10 | Fall from other furniture |
| W09 | ICD-10 | Fall on and from playground equipment |
| W10 | ICD-10 | Fall on and from stairs and steps |
| W11 | ICD-10 | Fall on and from ladder |
| W12 | ICD-10 | Fall on and from scaffolding |
| W13 | ICD-10 | Fall from, out of or through building or structure |
| W14 | ICD-10 | Fall from tree |
| W15 | ICD-10 | Fall from cliff |
| W16 | ICD-10 | Fall, jump or diving into water |
| W17 | ICD-10 | Other fall from one level to another |
| W18 | ICD-10 | Other slipping, tripping and stumbling and falls |
| W19 | ICD-10 | Unspecified fall |
| E880 | ICD-9 | Accidental fall on or from stairs or steps |
| E881 | ICD-9 | Accidental fall on or from ladders or scaffolding |
| E882 | ICD-9 | Accidental fall from or out of building or other structure |
| E883 | ICD-9 | Accidental fall into hole or other opening in surface |
| E884 | ICD-9 | Other accidental falls from one level to another |
| E885 | ICD-9 | Accidental fall on same level from slipping tripping or stumbling |
| E886 | ICD-9 | Fall on same level from collision, pushing, or shoving, by or with other person |
| E887 | ICD-9 | Fracture, cause unspecified |
| E888 | ICD-9 | Other and unspecified fall |
| Z91.81 | ICD-10 | History of falling |
| 781.2 | ICD-9 | Abnormality of gait |
| R26.89 | ICD-10 | Other abnormalities of gait and mobility |
| R26.81 | ICD-10 | Unsteadiness on feet |
| R26.9 | ICD-10 | Unspecified abnormalities of gait and mobility |

ICD = International Classification of Diseases. ICD is the classification standard to share, report, and monitor disease from different medical systems in a consistent and standardized way between hospitals, regions, and countries.

### 3.1.2 Zip Code Tabulation Areas and Community-Based Variables

In addition to patient-based variables, we collected socioeconomic variables from 5-digit zip code tabulation areas. For zip code tabulation area-based variables, this study collected socioeconomic variables from eight counties: Jefferson, Kenosha, Milwaukee, Ozaukee, Racine, Walworth, Washington, and Waukesha, which included 126 zip codes (Appendix A). If a patient resided in one of eight counties, the patient's socioeconomic variables would be added into the calculation of zip code tabulation area-based variables. Each zip code was used to connect the community information to the 5-year estimate data from the Census American Community Survey. We summarized the socioeconomic variables of the white rate, median household income, college educated rate, and privately insured rate (Appendix B). The socioeconomic data is publicly available through the U.S. Census Bureau. The variables and the calculations of variables are as follows:

#### 3.1.2.1 The Percentage of Whites

To obtain the 126 zip-level-based percentage of whites using the American Community Survey (ACS), we downloaded the data from the United States Census Bureau's website, using the 2014–2018 5-Year Estimates Data Profile, and chose the ZIP Code Tabulation Area as the geography of interest. We selected the data field with "Race" and "White Alone" under the "Topic" button to show the overall number of whites under each zip code. Then, we also selected the total number of populations under each zip code. The zip-based percentage of whites is the number of whites divided by the total population under each zip code. Therefore, the defined white rate for each zip code is:

$$White\ rate = \frac{Total\ population\ from\ Census\ , White\ alone}{Total\ population\ from\ Census}$$

#### 3.1.2.2 The Percentage of College-Educated Residents

To obtain the 126 zip-level-based percentage of college-educated residents, we acquired data from the United States Census Bureau's website, selected the "Data Profiles" option under the ACS page, selected the 2014-2018 5-Year Estimates Data Profile, and chose the ZIP Code Tabulation Area as the geography of interest. We chose the "population 25 years and over with a bachelor’s degree or higher" as a numerator and selected the similar number of populations under each zip code as a denominator:

$$College\ educated\ Rate = \frac{Population\ from\ Census, Bachelor's\ degree\ or\ higher}{Total\ population\ from\ Census}$$

#### 3.1.2.3 A Tabulation Area's Median Household Income

To obtain the 126 zip-level-based household income, we visited the United States Census Bureau's website, selected the "Data Profiles" option under the ACS page, selected the 2014-2018 5-Year Estimates Data Profile, and chose the ZIP Code Tabulation Area as the geography of interest. We chose the "median household income" to be included in our analysis.

#### 3.1.2.4 The Percentage of the Population with Insurance

As the insurance status was unavailable through the American Community Survey, we estimated the insurance coverage using the data from electronic health records from the Clinical Translational Science Institute. Insurance status was classified into public, private, other, and uninsured based on the payer's information from the electronic health records. Insurance status was classified into public, private, other, and uninsured based on the payer's information from the database when the patient was registered during the most recent visit. To obtain the 126 zip-level-based percentage of insurance coverage, we estimated using the zip-level calculation. Firstly, we calculated the number of patients who visited the hospital during the study period from electronic health records; then, we summarized the number of patients who are publicly, privately, and non-insured separately for each zip code. The percentage is calculated by dividing the number of patients with each type of insurance by the total number of patients in each zip code.

$$Insurance\ rate(public) = \frac{population\ in\ medical\ records,\ \ with\ public\ insurance}{population\ in\ medical\ records}$$

$$Insurance\ rate(private) = \frac{population\ in\ medical\ records,\ \ with\ private\ insurance}{population\ in\ medical\ records}$$

$$Non - insured\ Rate = \frac{population\ in\ medical\ records, without\ insurance}{population\ in\ medical\ records}$$

### 3.1.2.5 Area Deprivation Index

The Area Deprivation Index (ADI) was based on a measure created by the Health Resources and Services Administration and developed by Dr. Amy Kind [152]. It allowed for rankings of neighborhoods by socio-economic disadvantage in a region of interest. ADI ranged on a scale of 0 to 100, from the most affluent to the most disadvantaged, and according to mixed factors including income, education, employment, and housing quality. ADI was used to inform socioeconomic status, health delivery, and policy conditions, especially for the most disadvantaged neighborhood groups.

The ADI is built using American Community Survey (ACS) five-year estimates. The 2018 ADI, for example, uses ACS data for 2018, which is a 5-year average of ACS data obtained from 2014 to 2018. All limitations of the source data will be carried over into the ADI; results are subject to the accuracy and errors contained in the American Community Survey data release. The choice of geographic units will also have an impact on the ADI value. Because the Census Block Group is the closest approximation to a "neighborhood" in the case of the ADI, it is the geographic unit of construction. Therefore, it is suggested that the Area Deprivation Index be used only along with the 2014–2018 5-Year Estimates Data Profile from the American Community Survey. The result and summary of population and categorizations of the Area Deprivation Index are available in Appendix C, which shows the 4 quartile gaps of the fall and non-fall cohort groups of patient distributions.

#### 3.1.2.6 Rural-Urban Continuum Codes

The 2013 Rural-Urban Continuum Codes [153] defined metropolitan counties by the population size of their metro area and nonmetropolitan counties by their degree of urbanization and proximity to a metro area. The official metro and non-metro categories of the Office of Management and Budget have been subdivided into three metro and six non-metro categories. One of the nine codes is assigned to each county in the United States, municipality in Puerto Rico, and Census Bureau-designated county-equivalent area of the Virgin Islands or other inhabited island territories of the United States. This scheme enables researchers to subdivide county data into finer residential groups other than metro and nonmetro, which is especially useful for analyzing trends in nonmetro areas that are related to population density and metro influence. The Rural-Urban Continuum Codes were created in 1974. Since then, they have been updated every decade (1983, 1993, 2003, and 2013) and were slightly revised in 1988. Because of the new methodology used in developing the 2000 metropolitan areas, the 2013 Rural-Urban Continuum Codes are not directly comparable with the codes prior to 2000. Details and a code map can be found in the documentation.

The Rural-Urban Continuum Codes (RUCC) reflect a classification scheme that distinguishes metropolitan counties by the population size of their metropolitan areas and non-metropolitan counties by the degree of urbanization and adjacency to a metropolitan area.

This study split RUCC into six categories of metropolitan and non-metropolitan counties. Metropolitan counties referred to counties in all metropolitan areas defined by the Office of Management and Budget as of February 2013; non-metropolitan counties included all non-metropolitan counties as well as completely unlisted rural areas. RUCC can be used to assess a patient's living environment and to inform rural and urban differences in relation to other social and economic variables. For the convenience of this study, we split the area into metropolitan and non-metropolitan areas according to the RUCC code.

### 3.1.2.7 Definition of Healthcare Utilization

The healthcare utilization rate (UR) is defined at the zip code tabulation area level. The utilization rate is defined as the number of unique patients that had a fall-related diagnosis in the electronic health system from one of the eight counties during the study period. The utilization rates are calculated using these formulas:

$$\mathrm{UR}_{\mathrm{fall}} = \frac{Total\ population\ with\ diagnosis\ of\ falls\ in\ medical\ records}{Total\ population\ in\ medical\ records}$$

$$\mathrm{UR}_{\mathrm{general}} = \frac{Total\ population\ without\ diagnosis\ of\ falls\ in\ medical\ records}{Total\ population\ in\ medical\ records}$$

The diagnosis of falls in medical records was determined from the presence of ICD codes in Appendix A: List of ICD-10 Codes for the Falls Group. The utilization rate in each zip code of the southeast Wisconsin region was assessed to determine the impacts of median income, white rate, college educated rate, and private insured rate. The utilization rate was calculated in the fall patient group and the non-fall patient group in order to make a comparison between the associations between utilization and socioeconomic variables.

### 3.1.3 Geographical and Statistical Analysis

The number of patients with a history of falls divided by the total population in each zip code block area is how this study defined a predictor variable called a utilization rate. The utilization rate for non-fall patients (UR-non-fall) was deemed to be the number of patients without a fall history divided by the total population in each zip code block area. Consequently, this study associated the utilization rate variable with socioeconomic variables to evaluate the effect of each socioeconomic factor. All patient-based variables were merged from the patient level to the zip-code tabulation area level in order to generate a geographical analysis.

For geographical analysis, we plotted the utilization rate of senior adult falls and the socioeconomic variables in 126 zip codes. The area with a darker color shows a higher utilization rate and a varying degree of socioeconomic status. The map plotting was completed using Microsoft Excel (2016) and the creation of map charts, in which zip code and utilization rate were provided to generate the map.

To further quantify the association between socioeconomic variables and utilization rates, this study completed a multiple linear regression analysis. The multiple linear regression used the utilization rate of falls and the utilization rate of other care as predictor variables, employing socioeconomic analysis as an independent variable. This study plotted scatter plots and calculated the coefficient for each socioeconomic variable. Finally, multiple regression analysis was used to analyze the collective impact of social determinant factors on healthcare utilization rates.

All statistical analyses were performed using the R programming language. Statistical tests were two-sided, and the alpha was set at 0.05. This study calculated P-values using chi-square tests for categorical variables. Within table 1 of the comparison for two cohorts, this study used the odds ratio (OR) to measure the association with patient characteristics. ORs were calculated through a two-by-two contingency table. The table compares the size of the effect between the fall history group and the non-fall history group. Concerning each patient characteristic, an OR value larger than 1 indicated that patients with the corresponding characteristics were more likely to experience falls and visit the hospital, while an OR value smaller than 1 indicated that patients with the corresponding characteristics were less likely to experience falls. The 95% confidence interval demonstrated the 95% likelihood range of the OR based upon a normal distribution. A P-value of less than 0.05 indicated the difference in patient characteristics between the two groups to be statistically significant.

### 3.1.4 An Example and Calculation of Odds Ratios

Here we show an example of an odds ratio calculation. We use the odds ratio and the 95% confidence intervals to evaluate the association between the socioeconomic variables and fall adoptions. The odds ratio is the odds of success in the treatment group relative to the odds of success in the control group. This method is used in cases where the data is binary.

In a typical scenario where an odds ratio is used to evaluate the intervention effect, e.g., in a clinical trial, a two-by-two table can be presented as follows:

Table 2. two-by-two table showing associations between intervention and fall utilization

| Treatment | | With falls | Without falls |
|---|---|---|---|
| Intervention Group | | a | b |
| Control Group | | c | d |

The odds ratio (OR) and confidence interval can be calculated using the following formulas:

$$Odds\ ratio = \frac{a \times d}{b \times c}$$

$$Lower\ 95\%\ CI = e^{\left(\ln OR - 1.96\left(\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)\right)}$$

$$Upper\ 95\%\ CI = e^{\left(\ln OR + 1.96\left(\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\ \right)\right)}$$

Where $a$ is the number of treatment group with falls, $b$ is the number of treatment group without falls, $c$ is the control group with falls, and $d$ is the control group without falls. OR is the odds ratio calculated from the first formula, and 1.96 is the approximate z-value of the 95-percentile point of the standard normal distribution. In this example, the odds ratio shows the intervention effect. Therefore, the odds ratio demonstrates that patient in the intervention group have a $\frac{a\times d}{b\times c}$ likelihood of fall, compared with control group. The 95% confidence interval of the odds ratio ranges from $e^{\left(\ln OR - 1.96\left(\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)\right)}$ to $e^{\left(\ln OR + 1.96\left(\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\ \right)\right)}$.

For an effect from binary variables, when we use one of the socioeconomic variables to evaluate the socioeconomic factors, we can consider the target variable as a form of intervention. Therefore, we can provide statistical interpretations regarding the associations between target variables and fall utilization. For example, if we choose ‘female’ as a target binary variable, the two-by-two table can show as follows:

Table 3. A two-by-two table showing associations between gender and fall utilization for binary variables

| Sex | With falls | Without falls |
|---|---|---|
| Female | 38739 | 225518 |
| Male | 27617 | 203086 |

Therefore, the odds ratio calculation is as follows:

$$Odds\ ratio = \frac{38739 \times 203086}{27617 \times 225518} = 1.26$$

$$Lower\ 95\%\ CI = e^{\left(\ln 1.26 - 1.96\left(\sqrt{\frac{1}{38739}+\frac{1}{225518}+\frac{1}{27617}+\frac{1}{203086}}\right)\right)} = 1.24$$

$$Upper\ 95\%\ CI = e^{\left(\ln 1.26 + 1.96\left(\sqrt{\frac{1}{38739}+\frac{1}{225518}+\frac{1}{27617}+\frac{1}{203086}}\right)\right)} = 1.28$$

Under a normal distribution, the p-value of the odds ratio is  0.001. Therefore, it is demonstrated that gender is significantly associated with fall utilization based on the provided data. We state that female patients were 1.26 times more likely than males to have used fall healthcare adoption compared to male patients. Similarly, we can change the effect of gender to other binary socioeconomic variables to evaluate the associations between fall utilization and other binary socioeconomic variables.

From a statistical perspective, we conclude that females are 1.26 times more likely to utilize healthcare services after a fall (odds ratio [OR]: 1.26, 95% confidence interval [95% CI]: 1.24–1.28). Other categorical variables, including age groups, race, ethnicity, type of insurance, employment, language assistance, area deprivation index, and rural-urban continuum codes, can be calculated using the same formula and methodologies.

For an effect from category variables, we can consider the target as one of our most interested fields as an intervention group. The subject of interest was considered an individual group in the two-by-two table, and the rest of the subjects were summarized into another group. The two-by-two table can be shown as follows:

Table 4. A two-by-two table showing associations between gender and fall utilization for categorial variables

| Employment Status | With falls | Without falls |
|---|---|---|
| Retired | 42474 | 176855 |
| All other non-retired (including full time, part time, self-employed, not employed and disabled. ) | 23883 | 251829 |

Therefore, the odds ratio calculation is as follows:

$$Odds\ ratio = \frac{42474 \times 251829}{23883 \times 176855} = 2.53$$

$$Lower\ 95\%\ CI = e^{\left(\ln 2.53 - 1.96\left(\sqrt{\frac{1}{42474}+\frac{1}{251828}+\frac{1}{23883}+\frac{1}{176855}}\right)\right)} = 2.49$$

$$Upper\ 95\%\ CI = e^{\left(\ln 2.53 + 1.96\left(\sqrt{\frac{1}{42474}+\frac{1}{251828}+\frac{1}{23883}+\frac{1}{176855}}\right)\right)} = 2.58$$

Under a normal distribution, the p-value of the odds ratio is 0.001. Therefore, it is demonstrated that retirement is significantly associated with fall utilization based on the provided data. We show that retired patients were 2.53 times more likely than males to have used fall healthcare adoption compared to male patients. The 95% confidence interval is (2.49, 2.58), meaning that in 95% of cases, the estimation will generate an odds ratio between these ranges. Similarly, we can change the effect of employment status to other categorical socioeconomic variables to evaluate the associations between fall utilization and other socioeconomic variables.

### 3.1.5 Creation of the Scatter Plot

We used a scatter plot to show the associations between utilization rate and four socioeconomic variables. The scatter plots are based on the utilization of fall patient groups and the utilization of non-fall patient groups. For each group, we associated the utilization rate with median household income, College educated rate, White ratio, and private insurance rate, which shows the regression analysis for utilization rate and socioeconomic variables across 126 zip code tabulation areas of the Southeast Wisconsin area. The R-squared value and p-value are calculated from a single linear regression using the utilization rate as a dependent variable and socioeconomic variables as predictors, using the total sum of squares.

### 3.1.6 A Calculation of Multiple Linear Regression Results

In addition, we use multiple linear regression to quantify the socioeconomic effect on the utilization rate. A multiple linear regression result uses a quantifiable number to interpret the effect of socioeconomic variable changes. For example, when including income as a predictor in the multiple linear regression model, each $1,000 increase in median household income will be associated with a certain percentage increase in utilization rate. Such quantifiable numbers would be an effective way to conclude the impact of socioeconomic variables on healthcare utilization.

We use the zip-based representation of utilization rates for fall patients and the zip-based representation of utilization rates for non-fall patients. As we call the multiple linear regression model, the model shows the statistical summaries on fall and non-fall patient cohorts separately:

For patient with falls, the generated estimated are shown as follows:

```
       Call:
       lm(formula = df$falling_UR_rate ~ df$income +df$private_insured_rate + 
df$white + df$college_educated_rate)
       Residuals:
             Min        1Q    Median        3Q       Max 
       -0.041383 -0.005701 -0.000820  0.004356  0.058161 
       Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
       (Intercept)               1.971e-02  4.567e-03   4.315 3.28e-05 ***
       df$income                 2.948e-07  9.132e-08  -3.228   0.0016 ** 
       df$private_insured_rate   1.132e-01  6.331e-03  17.881  < 2e-16 ***
       df$white                 -1.411e-02  6.768e-03  -2.085   0.0392 *  
       df$college_educated_rate  1.748e-02  7.898e-03   1.766   0.0399 *  
       ---
       Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
       Residual standard error: 0.01225 on 121 degrees of freedom
       Multiple R-squared:  0.7579,
       F-statistic:  94.7 on 4 and 121 DF,  p-value: < 2.2e-16
```

For patient without falls:

```
       Call:
       lm(formula = df$general_UR_rate ~ df$income + df$private_insured_rate + 
           df$white + df$college_educated_rate)
       Residuals:
            Min       1Q   Median       3Q      Max 
       -0.11621 -0.03726 -0.00978  0.01784  0.72898 
```

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.358e-01  3.263e-02   4.162 5.94e-05 ***
df$income                7.883e-07  6.525e-07   1.208   0.2294
df$private_insured_rate  5.448e-01  4.524e-02  12.042  < 2e-16 ***
df$white                -1.075e-01  4.836e-02  -2.223   0.0281 *
df$college_educated_rate  1.597e-01  7.073e-02  -2.258   0.0257 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.08754 on 121 degrees of freedom
Multiple R-squared:  0.5926,
F-statistic:    44 on 4 and 121 DF,  p-value: < 2.2e-16
```

Therefore, we can summarize the results in Table 4 to present the multiple regression modeling of the impact of socioeconomic variables on falls and non-fall patients within the Southeastern Wisconsin area.

### 3.2 Results

Table 3 highlights that 495,041 patients visited the hospital between March 1st, 2020, and March 1st, 2022. In Figure 1, this study visualizes the association between the utilization gap and socioeconomic variables in a forest plot. Among all patients, 66,357 (13.4%) were diagnosed with fall-related conditions. Patients with a history of falls were typically older compared with those without a history of falls (79.4 ± 12.1 vs. 75.4 ± 11.6 years old). Patients in an older age group had an increasing risk of falling. It was also observed that patients who were retired were 2.53 times more likely to experience falls and access healthcare facilities.

The most important factor influencing healthcare utilization is age. Patients between the ages of 60 and 64 are 0.58 times more likely than older patients to visit a healthcare facility. As a patient's age increases, so does their utilization. Patients over the age of 85 have a 1.93-times higher utilization rate than younger patients.

The utilization of healthcare services is related to demographic factors. Within patients with a history of falls, the ratio of females was significantly higher in comparison to the ratio of females in patients without a history of falls. Females were 1.26 times more likely than males to have used fall healthcare services, according to OR datasets. White patients were 1.35 times more likely to utilize healthcare services compared with other racial groups, resulting in a higher proportion of falls. Asian and black races had significantly lower utilization of hospital-based care. In terms of ethnicity, the non-Hispanic population was 3.22-fold more likely to utilize such services compared to the Hispanic population, highlighting that Hispanic groups are underserved in the healthcare system.

Economic factors were also linked to the utilization of healthcare services. Patients with public insurance were 2.88 times more likely to use the services than those with other insurers. In addition, patients who were uninsured were less likely to use healthcare services. The OR value of 0.51 demonstrated that uninsured patients are an underserved population in the healthcare system. The uninsured status of patients is an indicator of lower socioeconomic conditions. There was no association between the ADI and patient access to hospital-based care.

The factor of location was also associated with a patient's access to health care. The hospital is situated in Wauwatosa, Wisconsin, which is in the suburban area of southeastern Wisconsin and is part of the great Milwaukee metropolitan area. Over 90% of patients in this study currently reside in the Milwaukee metropolitan area. Patients who live in metropolitan areas were 1.33-fold more likely to use the services in comparison to those living outside the metropolitan Milwaukee area, which shows that patients who live in rural areas are not likely to receive equal access to healthcare.

### 3.2.1 Geographic Map-Based Analysis

Fig. 2 shows the utilization rate of general and fall-based healthcare in the southeastern Wisconsin area to help understand the geographic distribution of utilization rates. The utilization rate map shows that utilization was significantly higher in the hospital's surrounding area. The utilization rate was lower in suburban and rural areas, which are situated far away from the hospital. To understand the socioeconomic variables, this study also plotted the visualization of socioeconomic variables in Southeast Wisconsin in Fig. 3, highlighting the college education rate, the white rate, income, and the private insurance rate. This study also observed significant gaps in differing counties in college education rates, racial distribution, and median household income levels. According to the map analysis, the Northwest area had both a high utilization rate and a high private insurance rate, indicating the possibility of correlations between the utilization rate and private insurance variables.

Table 5. Overall patient characteristics, socioeconomic, and demographic variable comparisons for patients with or without fall history.

|  | Patient with History of Falls |  | Patient without History of Falls |  | OR | 95% CI | p-value |
|---|---|---|---|---|---|---|---|
| Number of Patients | 66357 |  | 428684 |  |  |  |  |
| Age |  |  |  |  |  |  |  |
| Median, SD | 79.4 ± 12.1 |  | 75.4 ± 11.6 |  |  |  |  |
| 60 - 64 years old | 7626 | 11.5% | 78797 | 18.4% | 0.58 | (0.56, 0.59) | <0.001 |
| 65 - 69 years old | 9879 | 14.9% | 88288 | 20.6% | 0.67 | (0.66, 0.69) | <0.001 |
| 70 - 74 years old | 9686 | 14.6% | 74970 | 17.5% | 0.81 | (0.79, 0.83) | <0.001 |
| 75 - 79 years old | 8255 | 12.4% | 53008 | 12.4% | 1.01 | (0.98, 1.03) | 0.585 |
| 80 - 84 years old | 7556 | 11.4% | 39304 | 9.2% | 1.27 | (1.24, 1.31) | <0.001 |
| >85 years old | 23355 | 35.2% | 94317 | 22.0% | 1.93 | (1.89, 1.96) | <0.001 |
| Sex |  |  |  |  |  |  |  |
| Female | 38739 | 58.4% | 225518 | 52.6% | 1.26 | (1.24, 1.28) | <0.001 |
| Male | 27617 | 41.6% | 203086 | 47.4% | 0.79 | (0.78, 0.81) | <0.001 |
| Race |  |  |  |  |  |  |  |
| White | 54755 | 82.5% | 333216 | 77.7% | 1.35 | (1.32, 1.38) | <0.001 |
| Black | 8219 | 12.4% | 59619 | 13.9% | 0.88 | (0.85, 0.9) | <0.001 |
| Asian | 465 | 0.7% | 4079 | 1.0% | 0.73 | (0.67, 0.81) | <0.001 |
| Other | 1301 | 2.0% | 8396 | 2.0% | 1.00 | (0.94, 1.06) | 0.972 |
| Unknown | 1617 | 2.4% | 23374 | 5.5% | 0.43 | (0.41, 0.46) | <0.001 |
| Ethnicity |  |  |  |  |  |  |  |
| Hispanic | 1196 | 1.8% | 9148 | 2.1% | 0.84 | (0.79, 0.89) | <0.001 |
| Non-Hispanic | 63680 | 96.0% | 377610 | 88.1% | 3.22 | (3.09, 3.35) | <0.001 |
| Unknown | 1481 | 2.2% | 41926 | 9.8% | 0.21 | (0.2, 0.22) | <0.001 |
| Type of Insurance |  |  |  |  |  |  |  |
| Private | 11317 | 17.1% | 143964 | 34.8% | 0.41 | (0.4, 0.42) | <0.001 |
| Public | 53455 | 80.8% | 253031 | 61.2% | 2.88 | (2.82, 2.93) | <0.001 |
| Other | 532 | 0.8% | 5639 | 1.4% | 0.61 | (0.55, 0.66) | <0.001 |
| Uninsured | 871 | 1.3% | 10955 | 2.6% | 0.51 | (0.47, 0.54) | <0.001 |
| Employment Status |  |  |  |  |  |  |  |
| Retired | 42474 | 77.6% | 176855 | 59.7% | 2.53 | (2.49, 2.58) | <0.001 |
| Full Time | 3937 | 7.2% | 68125 | 23.0% | 0.33 | (0.32, 0.35) | <0.001 |
| Part Time | 1157 | 2.1% | 12331 | 4.2% | 0.60 | (0.56, 0.64) | <0.001 |
| Self Employed | 953 | 1.7% | 11443 | 3.9% | 0.53 | (0.5, 0.57) | <0.001 |
| Not Employed | 2757 | 5.0% | 17223 | 5.8% | 1.04 | (0.99, 1.08) | 0.095 |
| Disabled | 3446 | 6.3% | 10493 | 3.5% | 2.18 | (2.1, 2.27) | <0.001 |
| Interpreter Needed? |  |  |  |  |  |  |  |
| N | 52935 | 98.7% | 266673 | 98.4% | 2.40 | (2.35, 2.44) | <0.001 |
| Y | 679 | 1.3% | 4374 | 1.6% | 1.00 | (0.92, 1.09) | 0.944 |
| Area Deprivation Index |  |  |  |  |  |  |  |
| (Most Affluent) 0 - 25 | 4487 | 9.5% | 31705 | 10.6% | 0.91 | (0.88, 0.94) | <0.001 |
| 25 - 50 | 18256 | 38.6% | 110056 | 36.6% | 1.10 | (1.08, 1.12) | <0.001 |
| 50 - 75 | 14340 | 30.3% | 95220 | 31.7% | 0.97 | (0.95, 0.98) | <0.001 |
| 75 - 100 | 10215 | 21.6% | 63511 | 21.1% | 1.05 | (1.02, 1.07) | <0.001 |
| Rural-Urban Continuum Codes |  |  |  |  |  |  |  |
| Metropolitan Area (> 1m) | 45483 | 90.7% | 264089 | 83.4% | 1.36 | (1.33, 1.38) | <0.001 |
| Metropolitan Area (250k - 1m) | 2321 | 4.6% | 21061 | 6.6% | 0.70 | (0.67, 0.73) | <0.001 |
| Metropolitan Area (< 250k) | 210 | 0.4% | 1905 | 0.6% | 0.71 | (0.62, 0.82) | <0.001 |
| Micropolitan Area (>20k) | 1022 | 2.0% | 15270 | 4.8% | 0.42 | (0.4, 0.45) | <0.001 |
| Small Town area (2.5k - 20k) | 533 | 1.1% | 7317 | 2.3% | 0.47 | (0.43, 0.51) | <0.001 |
| Rural area (< 2.5k) | 576 | 1.1% | 7093 | 2.2% | 0.52 | (0.48, 0.57) | <0.001 |

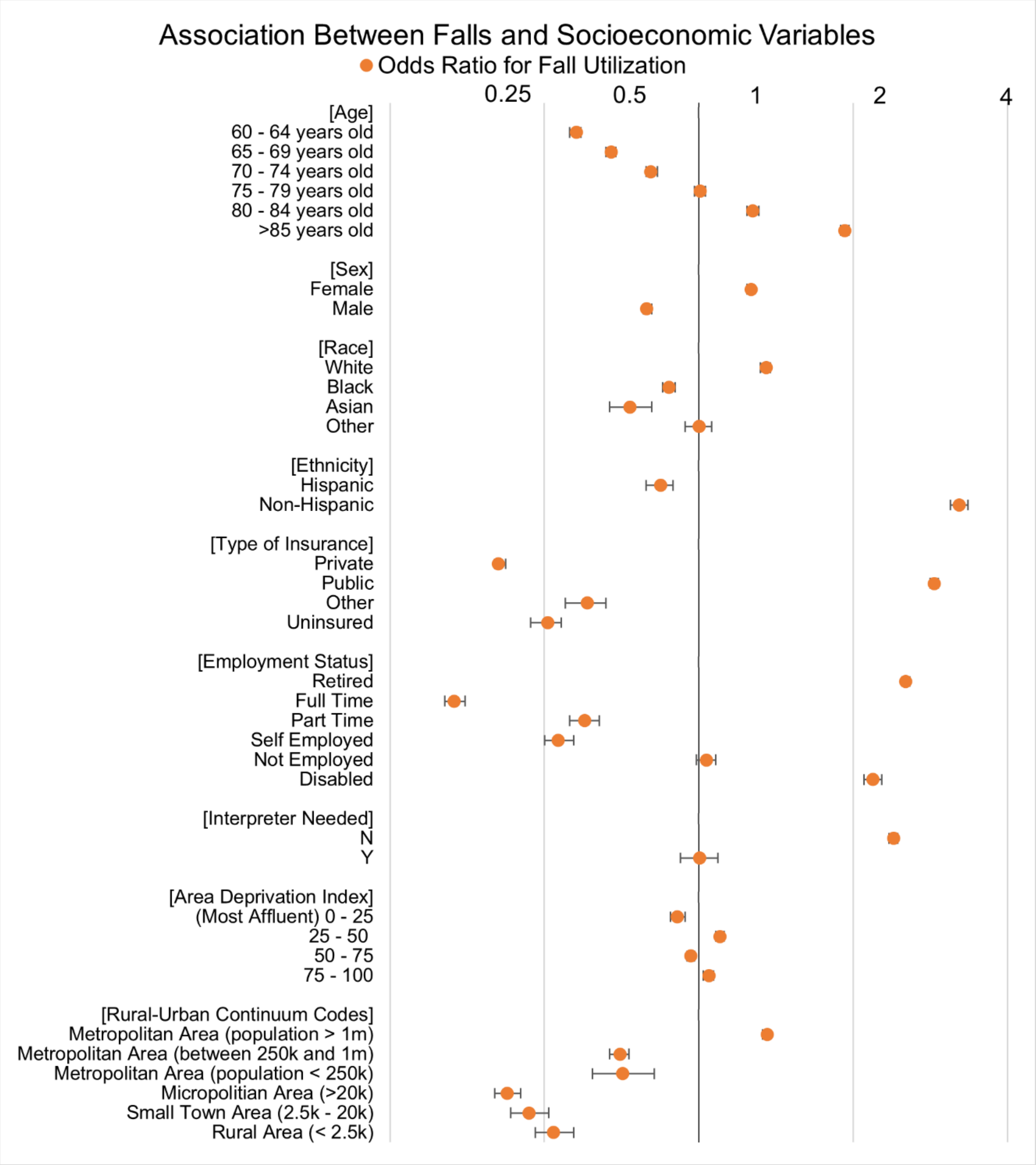


Fig. 1. forest plot of association between falls and socioeconomic variables

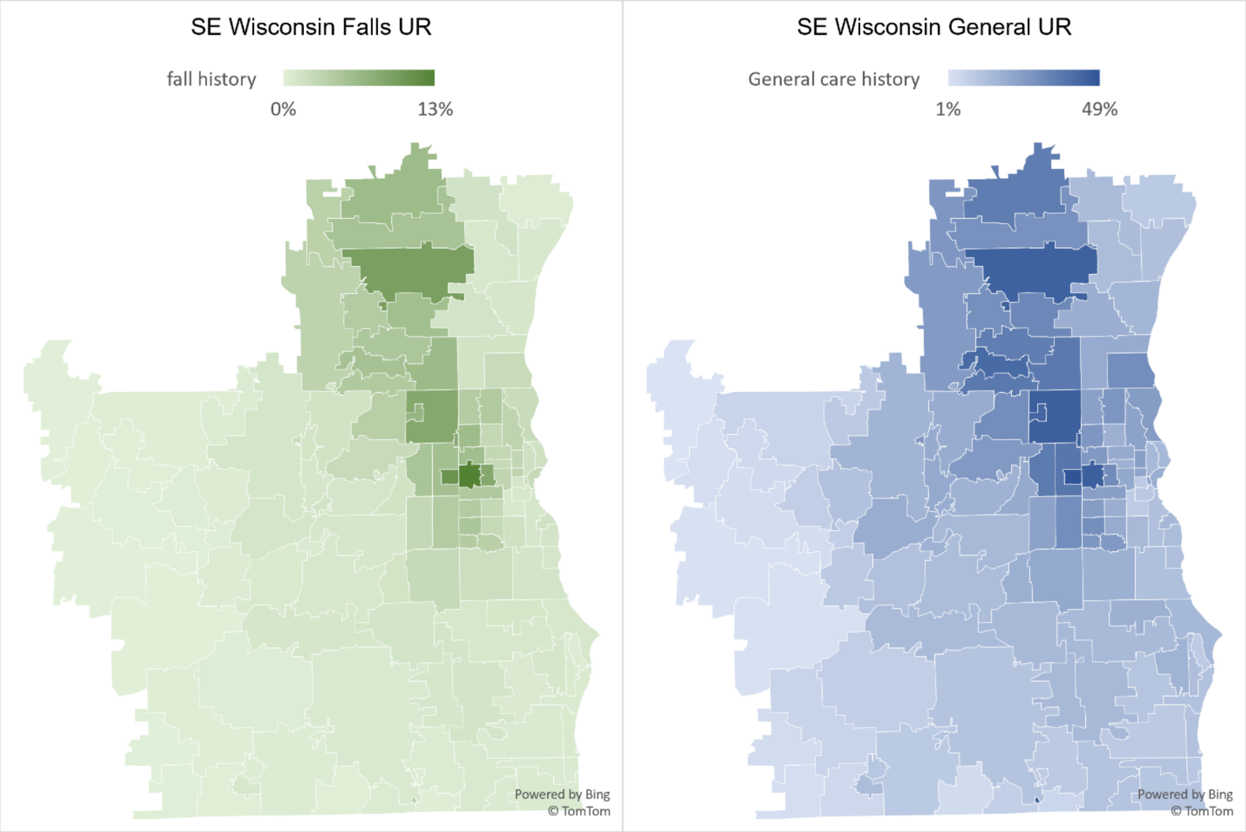


Fig. 2. utilization rate of general and fall-based care in southeastern Wisconsin

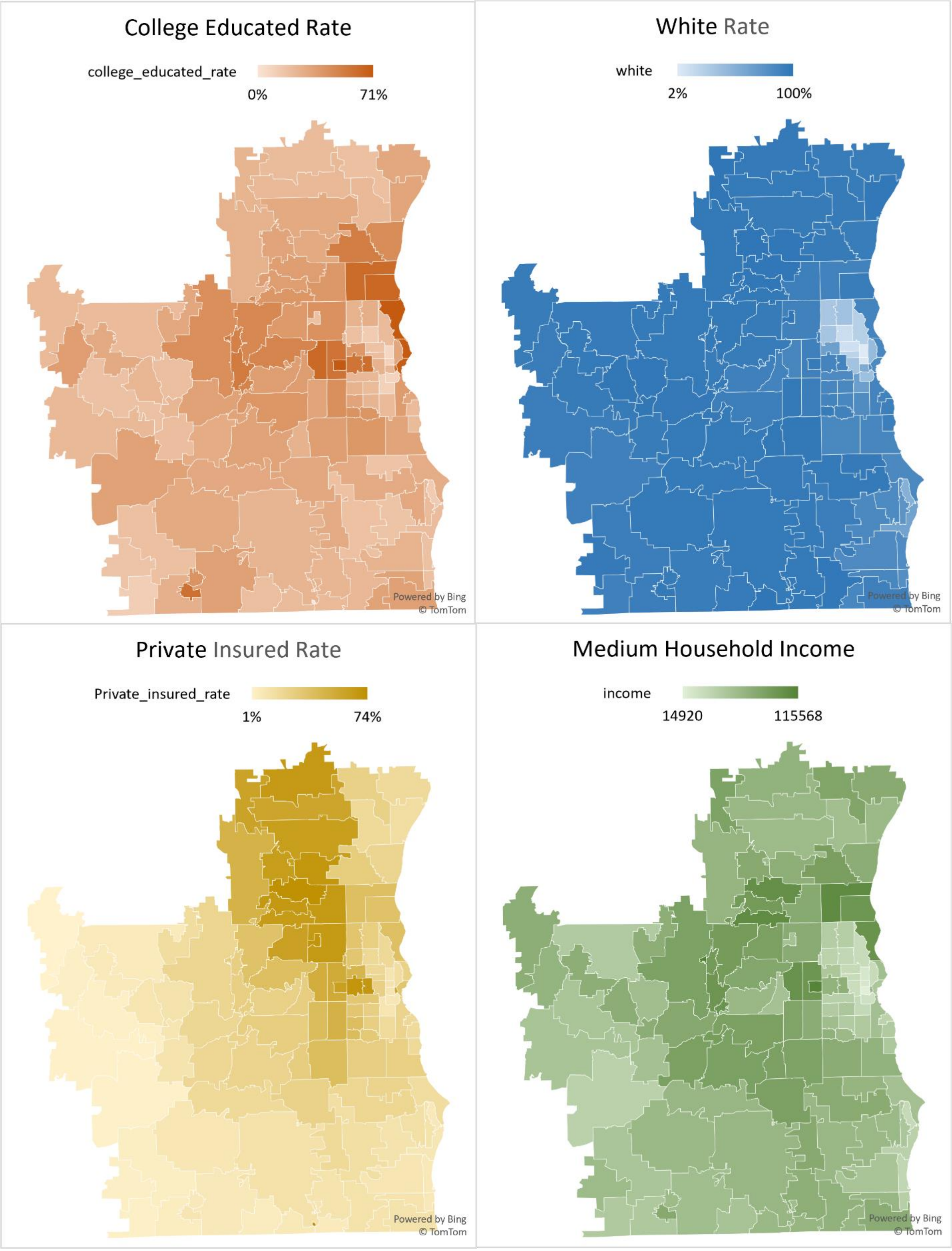


Fig. 3. visualization of four socioeconomic variables in southeast Wisconsin area in a zip code tabulated map.

### 3.2.2 Area-based Analysis: Linear Regression Assessment

Based on the linear regression analysis in Figure 4, it appears that the private insurance rate was the most robust predictor of healthcare utilization rates for both fall-based and non-fall-based healthcare, with R-squared values of 0.68 and 0.79, respectively. The college educated rate and median household income also showed a positive association with healthcare utilization rates.

From the multiple regression analysis in Table 4, the white ratio, college education rate, and private insurance rate were found to be associated with healthcare utilization for patients with a history of falls. Concerning patients without a history of falls, the college education rate and private insurance rate were positively associated with healthcare utilization.

These results suggest that socioeconomic status plays an important role in determining healthcare utilization rates, particularly for fall-related care. Patients with private insurance and higher education levels appear to have better access to healthcare services, while those from racial and ethnic minority groups, lower-income households, and uninsured individuals may experience barriers to accessing care.

Table 5. Multiple regression modelling of the impact of socioeconomic variables for falls and non-fall patients within Southeastern Wisconsin area

| Social Determinant factors | Coefficient Estimate | Standard Error | P-value |
|---|---|---|---|
| Patient of Fall History: | | | |
| Income | 2.948E-07 | 9.138E-08 | 0.0016 |
| **White Ratio** | **-0.01411** | **0.006768** | **0.0492** |
| **College educated rate** | **0.01748** | **0.007898** | **0.0399** |
| **Private insurance rate** | **0.1132** | **0.006331** | **<0.0001** |
| Patients of general care | | | |
| Income | 7.883E-07 | 6.525E-07 | 0.2294 |
| White Ratio | -0.1075 | 0.04836 | 0.0281 |
| **College educated Rate** | **0.1597** | **0.07073** | **0.0257** |
| **Private insurance rate** | **0.5448** | **0.04524** | **<0.0001** |

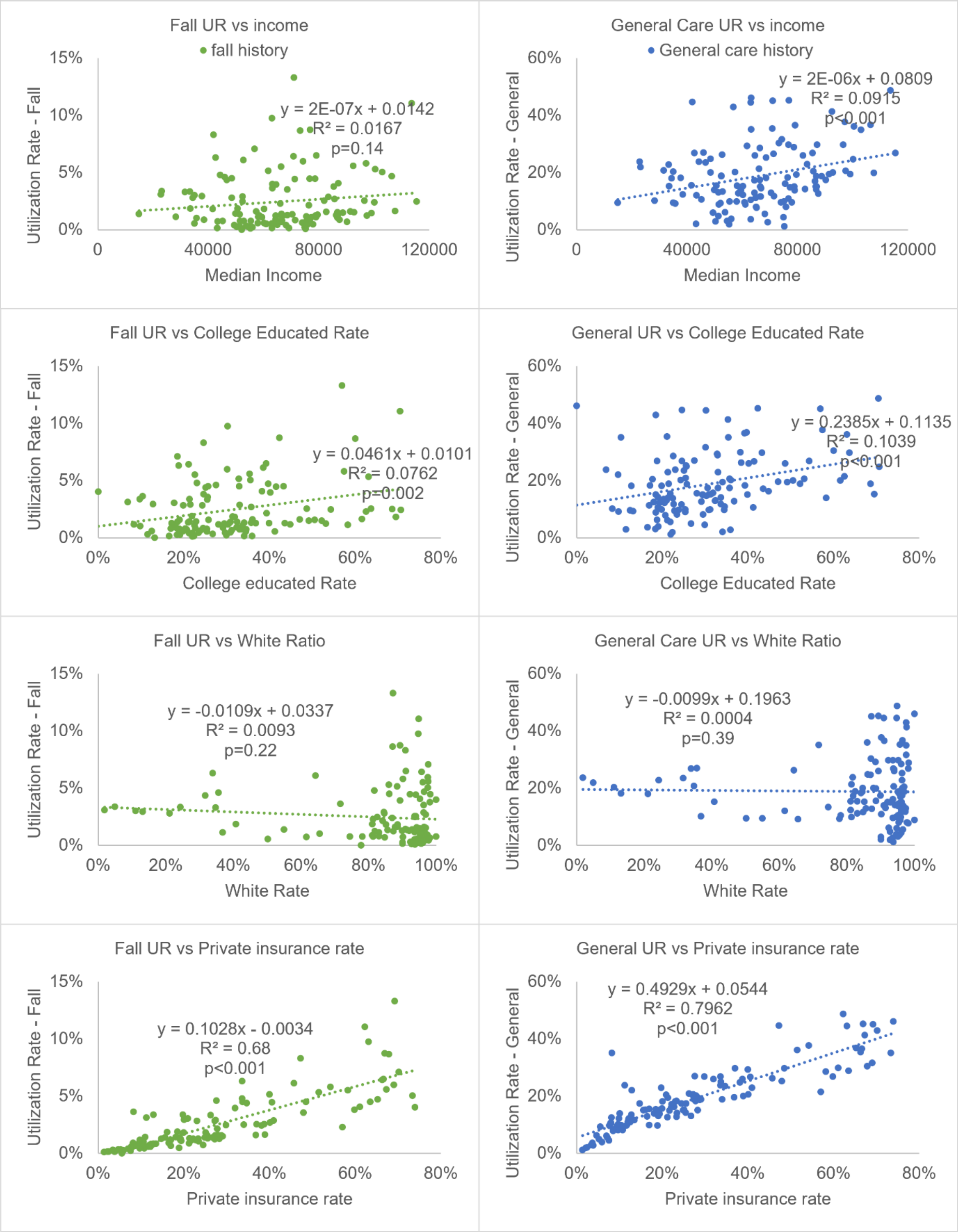


Fig. 4. scatter plot and regression analysis for utilization rate and socioeconomic variables across 126 zip code tabulation areas of the southeast Wisconsin area

### 3.3 Discussion

A majority of current healthcare outcome studies focus on risk factors for falls. However, there is a lack of studies on patient socioeconomic effects on healthcare utilization. The study reveals the potential socioeconomic inequities and disparities in the healthcare adoption of senior adults. With the growing percentage of senior adult populations, specific strategies are needed to address the disparities in adoption among underserved senior populations.

We completed a cohort study based on a single tertiary care center in Milwaukee, including 495,041 senior adults. This study's results clearly demonstrated that lower socioeconomic status is associated with lower utilization of healthcare services for general healthcare and fall-based care. Data revealed that healthcare services did not reach all patients equally, especially those with lower socioeconomic status. This included those who were uninsured, living in a low-income and low-education community, or living far away from the metropolitan area. The study results are consistent with previous studies on the association between demographic and socioeconomic factors and healthcare utilization. Many studies [31]–[33] have shown that older age, female gender, and certain racial and ethnic groups are more likely to utilize healthcare services, which is consistent with the findings of this study. Additionally, previous studies have found that individuals with lower socioeconomic status and uninsured individuals are less likely to utilize healthcare services [93], which is also consistent with the findings in this study. The geographical distribution of healthcare utilization has also been studied in previous research. Many studies show that access to healthcare services is often unevenly distributed [154], with urban areas having higher utilization rates than rural areas, which is again consistent with the findings of this study [93], [154]. Furthermore, socioeconomic factors such as median household income, college education, and private insurance rates were positively correlated with healthcare utilization, as demonstrated by regression analysis on each zip code tabulation area. These findings suggest that economic conditions play a critical role in access to healthcare services and healthcare utilization.

According to the U.S. Census Bureau, the US is facing a rapid aging trend [155]. There were a total of 76 million births in the United States from 1946 to 1964 [156]. The baby boomers are currently the most at risk for falling, and approximately 16.5% (54.1 million) of the total population reaches 65 years of age or older [19]. It is imperative to manage the risk of falls for the elderly. On the one hand, patient falls are associated with other ongoing clinical conditions, such as hypertension [157], physical or cognitive impairments [17], medication [158], and environmental hazards [93]. The analysis of elderly falls may reveal key correlations with other clinical risk factors. Additional studies can uncover the association between falls and various clinical factors. Leveraging these clinical factors can help reduce the fall condition systematically.

While clinical factors have been demonstrated to be the centerpiece of care [30], the topic of equitable healthcare has been overlooked. [159] Few studies have investigated the association between fall risks and socioeconomic status, which may widen the gap between patients of different socioeconomic statuses. With this study's findings, the authors believe it is critical to understand how social determinants of health can affect senior adults' access to healthcare. This study specifically emphasized the issue of social and economic disparities among patients in the Southeastern Wisconsin area. Since social determinant variables are publicly available in Census data, such an analysis can be easily replicated in other cities or counties in the United States. Furthermore, this study's analysis provided quantitative results to measure the effect of social gaps on the utilization of healthcare services. The quantitative results revealed the indispensable value of closing socioeconomic gaps, which can lead to the realization of equal healthcare.

This study demonstrated that senior adult patients from socially marginalized groups face underutilization of healthcare services (Table 3). This study further showed that the healthcare utilization gap is significantly associated with many socioeconomic determinant factors (Table 4, Figures 2-4). Without the collaboration of healthcare professionals, it is unlikely that the utilization gap will be closed automatically. Since offering equitable healthcare leads to more efficient healthcare systems and better health outcomes, it is essential to systematically develop policies and practices to balance equity and efficiency for equal care.

### 3.3.1 Gender Disparity

According to Table 3, female patients are 1.23-fold more likely to have a fall history than male patients, which is consistent with other studies. This higher risk may be due to biological gender differences, as a study suggests that increased gait variability during dual-task assignments may contribute to the higher risk of falling in women. [160] While environmental hazards that cause falls can differ between men and women, further research is required to identify the clinical risk factors that differ between the two genders.

Other studies support this observation, stating that the greater risk was associated with increased gait variability [161]. The increased gait variability in women during multi-tasking can contribute to their increased risk of falling and, thereby, to their known greater risk of fractures. Other studies also discussed the fact that the environmental hazards that cause falls in men and women can vary. Men were likely to fall due to a loss of support (of the floor when standing or of the chair when seated). Berg [43] found that women were more likely to trip or stumble. However, no research has been conducted on the clinical factors that differ between men and women and lead to such fall events. It is possible that the clinical risk of fall factors can also vary between men and women. This study suggests that additional studies are warranted to reveal the gender difference in clinical factors that drive falls.

By understanding the gender-specific clinical factors associated with falls, healthcare providers can develop targeted interventions and prevention strategies to reduce fall risks for both men and women. This approach can lead to more personalized care and improved outcomes for older adults, ultimately enhancing their overall health, well-being, and quality of life.

### 3.3.2 Insurance Disparity

Our study reveals that patients with public insurance have a 2.64-fold higher risk of falling than those with private insurance, emphasizing the significant influence of unequal healthcare access on fall risk. Elderly patients without private insurance are particularly susceptible to falls, and the healthcare utilization rate demonstrates that the distance to the care center substantially impacts access to healthcare services. This disparity in healthcare utilization is further exacerbated by economic factors.

Considering that private insurers typically have a higher adoption of healthcare services, the adjusted odds ratio (OR) could be biased (higher than 2.64), which strongly suggests that public insurance is an essential risk factor for falls. It is crucial to address these disparities in insurance coverage, as they have a considerable impact on fall risk and access to healthcare services.

Efforts to reduce the disparities in insurance coverage should include policy initiatives that aim to expand public insurance benefits or provide more affordable private insurance options to those in need. Additionally, collaboration between healthcare providers, community organizations, and policymakers is necessary to develop targeted interventions and resources to support at-risk populations. By addressing insurance disparities and working to ensure equitable access to healthcare services, it is possible to reduce the risk of falls among senior adults and ultimately improve the overall health and well-being of this vulnerable population.

### 3.3.3 Geographic Distribution Disparity

Patients residing in the Northeast and in closer proximity to the care center demonstrate a higher overall utilization rate compared to those living further away. The distance to the care center significantly impacts utilization rates, irrespective of fall history. Utilization rates reveal that approximately 5–10% of the total population has visited the care center for diagnosing or treating falls, while 20–50% of the total population has previously visited the care center for non-fall-related diagnoses. Patients living at a greater distance from the hospital area struggle to receive the same level of access as those residing closer to the hospital's vicinity. Utilization rates fluctuate depending on the distance to the care center, with patients living in metropolitan cities experiencing a 1.36-fold higher likelihood of receiving care.

This geographic distribution disparity highlights the importance of addressing accessibility issues in healthcare services, particularly for elderly patients at risk of falls. Limited access to healthcare facilities can lead to delayed diagnosis, treatment, and interventions, potentially exacerbating the consequences of falls for patients living in remote or rural areas. To mitigate these disparities, healthcare systems and policymakers should consider implementing solutions such as telemedicine services, mobile healthcare clinics, and community-based fall prevention programs. These strategies can help bridge the gap in healthcare access for elderly patients, regardless of their geographic location, ultimately improving the overall health outcomes for this vulnerable population.

### 3.3.4 Rural and Urban Healthcare Utilization Disparity

The healthcare utilization rates for the fall patient group and non-fall patient group are 90.7% and 83.4%, respectively. As demonstrated by the Rural-Urban Continuum Code variables, patients living in metropolitan areas with a population of 1 million or more are 1.33 times more likely to receive healthcare services compared to patients who do not live in the Milwaukee metropolitan area. This observation underscores the utilization disparity between rural and urban areas in Southeast Wisconsin.

These disparities in healthcare utilization between rural and urban populations can be attributed to several factors, including reduced availability of healthcare facilities, fewer specialized healthcare providers, and limited transportation options in rural areas. Additionally, rural populations often face socioeconomic challenges that further exacerbate these disparities, such as lower income levels, higher rates of uninsured individuals, and reduced access to education.

To address these rural and urban healthcare utilization disparities, targeted interventions and policy changes are needed. Expanding the availability of healthcare services in rural areas by establishing satellite clinics, incentivizing healthcare providers to work in underserved communities, and utilizing telehealth services can help improve access to care for rural residents. Furthermore, implementing educational initiatives and community-based programs to raise awareness about fall prevention and management among rural populations can contribute to better health outcomes for elderly patients at risk of falls.

### 3.3.5 Racial Disparity

Compared with other racial minorities, patients who are white or Caucasian have a 1.35-times higher likelihood of receiving healthcare services. This disparity can be attributed to various factors, including differences in socioeconomic status.

Socioeconomic factors, such as income, education, and employment opportunities, often play a significant role in determining an individual's access to healthcare services. In many cases, racial minorities face unique challenges that may limit their access to care, such as language barriers, cultural differences, and experiences of discrimination or bias within the healthcare system. Furthermore, racial minorities are more likely to be uninsured or underinsured, which can pose additional barriers to accessing healthcare services.

In addition to socioeconomic factors, residential segregation and geographic disparities may also contribute to racial disparities in healthcare utilization. Racial minorities are often more likely to live in areas with limited access to healthcare facilities or in communities with fewer resources dedicated to healthcare services. This can result in decreased availability and quality of care for these populations.

To address and reduce racial disparities in healthcare utilization, targeted interventions and policy changes are necessary. Strategies to consider include increasing cultural competency training for healthcare providers, implementing language assistance programs, and promoting healthcare policies that address the unique needs of diverse populations. Additionally, investing in community-based initiatives and outreach programs can help raise awareness about health issues and encourage preventive care among racial minority populations.

### 3.3.6 The Quantitative Effects of Socioeconomic Variables on Utilization Rate

This study hypothesized that economic conditions play a critical role in accessing and utilizing healthcare services. Due to privacy concerns, this study was unable to directly use patient income data. Instead, it conducted a regression analysis on each zip code tabulation area and found that healthcare utilization was positively correlated with socioeconomic factors such as median household income, college education, and private insurance rates. Since these factors are highly correlated with patient income, it can be reasonably inferred that economic conditions play an essential role in the healthcare system. For example, a college education is a predictor of higher income and a better understanding of routine care, leading to more frequent hospital visits.

In multiple regression analysis, this study quantified the effect of socioeconomic variables on healthcare utilization. For instance, a 1% increase in college education was associated with a 0.017% increase in falls care utilization, and a 1% increase in private insurance rate corresponded to a 0.113% increase in falls care utilization (Table 4). It is essential to note that different cities have varying socioeconomic conditions and healthcare services, so identical interpretations cannot apply to all areas. However, this methodology can be replicated in other cities or states to assess the relationships between healthcare service utilization and socioeconomic variables, as well as the social determinants of health.

Although this study could not identify the underlying driving factors, the analysis demonstrated how socioeconomic factors are associated with healthcare adoption and inequalities, which are closely related to social determinant factors. Such analysis can help assess the social determinants of health for a patient cohort in a specific area, providing insights into how social and economic factors can affect healthcare visits. This analysis suggests that economic factors have a significant impact on access to healthcare services and the use of fall care, underscoring the importance of promoting policies that ensure more equitable care for all patients, regardless of their social standing.

Furthermore, it is crucial to consider that the study's findings can provide a basis for designing targeted interventions and policies. By understanding the relationship between socioeconomic variables and healthcare utilization, healthcare providers and policymakers can develop programs that address the specific needs of communities with lower socioeconomic status.

Such interventions may include increased investment in healthcare infrastructure in underserved areas, financial support for low-income patients to access healthcare services, and community-based health education and outreach initiatives.

### 3.4 Limitations of the Study

There were several potential limitations to this study that should be considered. Firstly, as with many social determinant studies, individual patient income information was not collected, and the median income in each ZIP code was used as a proxy for patient income, which could introduce potential bias. Secondly, the patient cohort in this study represented a regional hospital system in Milwaukee, which has one of the highest racial segregation scores in the US [162]. Therefore, the patient characteristics may not be generalizable to other healthcare systems and areas. Thirdly, this study only focused on a few key social determinant factors, and there are other potential factors, such as environmental and cultural factors, that could be considered in future studies. The goal of this study was to identify potential gaps in the adoption of telemedicine and facilitate future research in this area. Fourthly, the pandemic may have had a wide range of effects on different types of health services, and this study only measured telemedicine adoption. Other health services such as cancer treatment, chronic condition management, laboratory services, and pharmacy services were not included in this study, which could potentially distort results. Finally, this study did not include clinical conditions or issues, suggesting that older patients may have used more healthcare services simply because they had a higher level of illness and required additional services. Future studies must investigate how to analyze and integrate cultural-based variables to achieve equal access to healthcare services from a variety of perspectives.

### 3.5 Future Work

Future work could be focused on the investigation of the healthcare disparity in other areas of the United States. A multi-center study of the disparity could ensure the study is generalizable and becomes a systematic methodology to discover health disparity issues. Additionally, interventions aimed at reducing healthcare disparities in fall utilization should be explored, such as targeted education and outreach to underserved communities, improved access to preventative services, and cultural competency training for healthcare providers. By addressing the root causes of healthcare disparities in fall utilization, we can work towards creating a more equitable healthcare system that ensures all individuals receive the care they need to maintain their health and wellbeing.

It is essential to note that different cities have varying socioeconomic conditions and healthcare services, so identical interpretations cannot apply to all areas. However, this methodology can be replicated in other cities or states to assess the relationships between healthcare service utilization and socioeconomic variables, as well as the social determinants of health.

Additionally, future research should explore the potential effects of other socioeconomic factors not included in this study, such as employment status, family structure, and neighborhood characteristics. These factors may also contribute to disparities in healthcare access and utilization. Longitudinal studies could be conducted to examine the causal relationships between these factors and healthcare utilization over time.

### 3.6 Conclusions

The study conducted a cohort study on 495,041 senior adults from a single tertiary care center in Milwaukee to determine the association between lower socioeconomic status and healthcare utilization for general healthcare and fall-based care. The results showed that healthcare services did not reach all patients equally, especially those with lower socioeconomic status. Many previous studies have shown that individuals with lower socioeconomic status and uninsured individuals are less likely to utilize healthcare services. Additionally, the geographical distribution of healthcare utilization has also been studied in previous research, with many studies showing that access to healthcare services is often unevenly distributed. The study emphasized the issue of social and economic disparities among patients in the Southeastern Wisconsin area and provided quantitative results to measure the effect of social gaps on the utilization of healthcare services. The study also showed that patients from socially marginalized groups face underutilization of healthcare services, and the healthcare utilization gap is significantly associated with many socioeconomic determinant factors. Gender differences and insurance disparities were also observed. The study suggests that additional studies are warranted to reveal the gender difference in clinical factors that drive falls, and public insurance can be an important risk factor for falls. Patients who live in the Northeast and are close to the care center have a higher overall utilization rate of healthcare services than those who live farther away.

# 4. Study Two: Interpretable Machine Learning Text Classification for Clinical Computed Tomography Reports – A Case Study of Temporal Bone Fracture

Abstract:

Machine learning has demonstrated remarkable success in numerous applications, including the classification of patients' diagnostic outcomes based on free-text clinical notes. However, the complexity of machine learning models often makes it challenging to interpret the mechanisms behind their classification results. In the second study, we investigated interpretable representations of text-based machine learning classification models, focusing on temporal bone fractures in computed tomography (CT) text reports. In this study, we created machine learning models to classify temporal bone fractures based on 164 temporal bone CT text reports. We adopted four well-known algorithms: XGBoost, Support Vector Machine, Logistic Regression, and Random Forest.

To interpret the models, we used two major methodologies:

(1) Word Frequency Score (WFS): We calculated the average word frequency score for keywords, which represents the frequency gap between positive and negative classified cases. This helps learn the differences in keyword usage between fracture and non-fracture cases.

(2) Local Interpretable Model-Agnostic Explanations (LIME): We used LIME to show the word-level contribution to bone fracture classification. LIME helps in visualizing the contribution of specific keywords to the classification results.

In the temporal bone fracture classification, the random forest model achieved an average F1-score of 0.93, indicating high classification performance. The WFS analysis revealed a difference in keyword usage between fracture and non-fracture cases, providing insights into the critical terms that distinguish the two categories. Additionally, LIME visualizations showed the keywords'

contributions to the classification results, making it easier for physicians to understand the machine learning predictions. The evaluation showed the highest interpreting accuracy of 0.97, which signifies a high level of transparency in the classification process.

The interpretable text explainer developed in this study can improve physicians' understanding of machine learning predictions and increase their trust in computerized models. By providing simple visualizations, our model can support more transparent computerized decision-making in clinical settings.

Future research can extend this work by investigating other interpretable methods and applying them to different clinical text datasets, improving the generalizability of our findings. Furthermore, studies can focus on integrating interpretable text classifiers into real-world clinical workflows, allowing physicians to make better-informed decisions based on the insights provided by machine learning models. Ultimately, this can lead to improved patient care and a more efficient use of healthcare resources.

## 4.1 Methodology

The second study used machine learning models to classify fractures based on text reports and two methodologies for interpretation, resulting in high interpretation accuracy. We also used an interpretable machine learning framework to visualize the importance of word factors in the final classification result. This study can help physicians use technology to make more informed decisions as well as increase trust in computerized models. Fig. 5 is a graphical abstract of this study. We first start by collecting data, followed by word frequency analysis, a text classification model, and using a text explainer to interpret the word-level factors in the classification result.

We collected a set of 164 clinical temporal bone CT reports from the Clinical Research Data Warehouse of the Clinical and Translational Science Institute of Southeastern Wisconsin. We first created a vector representation of CT reports [37] and built text classification models. A follow-up classification performance was evaluated. To explain the machine learning model, we provided two types of model interpretation. The first type is text feature analysis, which generates feature importance scores as well as word frequency scores; the second type is a text explainer using LIME [38], which provides a variety of interpretations of the classification results.

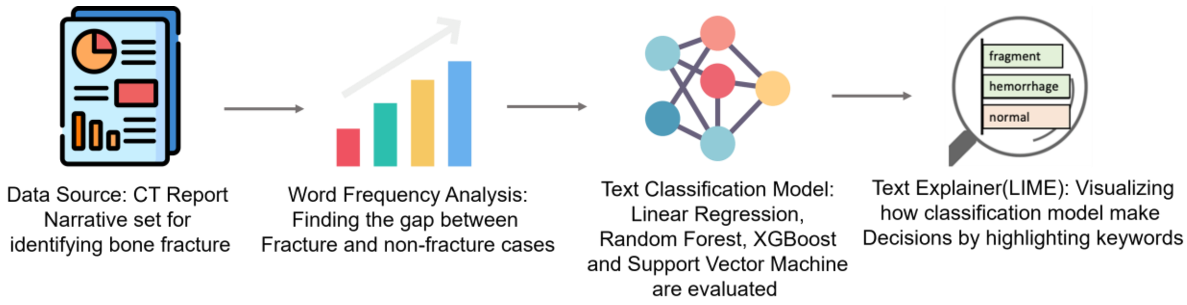

fragment
hemorrhage
normal
Data Source: CT Report Narrative set for identifying bone fracture
Word Frequency Analysis: Finding the gap between Fracture and non-fracture cases
Text Classification Model: Linear Regression, Random Forest, XGBoost and Support Vector Machine are evaluated
Text Explainer(LIME): Visualizing how classification model make Decisions by highlighting keywords

Fig. 5. Overview of our study. we first used CT text reports to construct a text-based classification model.

### 4.1.1 Data Source

The data source for this study was obtained through a request submitted to the Froedtert Health System i2b2 cohort query tool. The i2b2 tool aids in integrating genomic and clinical data from healthcare institutions and is maintained by the Clinical Research Data Warehouse (CRDW) of the Clinical & Translational Science Institute (CTSI) of Southeastern Wisconsin [39]. To define a manageable patient cohort for the text analysis, specific diagnosis codes were chosen. These codes helped to identify patients most likely to have a temporal bone fracture confirmed by a radiologist, thus being considered "clinically abnormal."

Upon submitting the query, an identified accession list of CT exams was generated for the study team. The team then collaborated with the business analyst of biomedical informatics to request a custom extraction of the imaging narratives and impressions from the data warehouse. These narratives and impressions were subsequently de-identified for integration with the text analysis of the study.

The query was further refined to include only adults aged 60–65, yielding a final normal cohort of 119 patients and a temporal bone fracture cohort of 45 patients. Each patient's narrative was included in only one clinical text, resulting in a total of 164 documents for the study. All documents have been submitted to the supplemental files of this study. Table 5 shows an example of the medical text sample used in the analysis:

Table 6. A Sample Computed Tomography Document Evaluation

| |
|---|
| Document #1:<br>1. Old comminuted fracture of the right middle cranial fossa with multiple bullet fragments lodged within it as described above. There is disruption/disolution of the right ossicular chain but the inner ear structures are intact. |

2. Adjacent residual right mastoid air cells are chronically opacified.

Examination reviewed by Dr. [NAME] and reported findings confirmed by Dr. [NAME].

Clinical Indication: Post traumatic right otalgia.

Techniques: 0.625 mm thick contiguous axial scans of temporal bones were acquired. Coronal reformats were generated and reviewed.

Comparison: None.

Findings: Again, visualized are severely comminuted old fractures of the right middle cranial fossa with multiple bullet fragments within the bones of the right skull base the right middle ear cavity and right anterior mastoid air cells. The roof of the right middle year cavity is dehiscent and there are dislocations/resorptions of components of the right ossicular chain. Only the body of the right incus is well visualized.

Multiple bullet fragments are lodged within the clivus sphenoid bone prevertebral soft tissues and in the infratemporal fossa. The residual mastoid air cells are opacified.

The left mastoid air cells appear well-aerated. The left middle ear cavity and ossicular chain are preserved. The left mastoid air cells appear unremarkable. Bilateral inner ear structures appear normal in morphology and density. Internal auditory canals appear symmetrical and normal in size bilaterally. Vestibular aqueducts are not dilated.

This example illustrates the type of clinical narratives and impressions used in the study, which the machine learning models were trained on to classify patients' diagnostic outcomes.

### 4.1.2 Text Pre-Processing

Text pre-processing is an essential part of natural language processing, a field of artificial intelligence (AI) that focuses on the interaction between computers and humans through natural language. The primary goal of NLP is to enable computers to understand, interpret, and generate human language in a way that is both meaningful and useful. Text pre-processing is an essential step in NLP, as it helps clean and structure raw text data, making it suitable for analysis and model training. In this study, we performed several text pre-processing steps on the clinical reports to ensure the text data was ready for further analysis. The pre-processing included five steps:

(1) Removing non-word elements: All non-word elements, such as numbers, punctuation, and special characters, were removed from the clinical reports. This was achieved using regular expressions. (2): Converting text to lowercase: All words in the clinical reports were converted to lowercase letters to maintain consistency and reduce redundancy in the dataset. (3): Removing stop words: Stop words, which are common words that do not carry significant meaning, were removed using the Natural Language Toolkit (NLTK) stop-word list. (4): Lemmatizing words: word lemmatization was performed to reduce words to their base or dictionary form, removing noun declination and verb conjugations. (5): Correcting spelling and acronyms: Any incorrect spellings and acronyms were corrected to ensure an accurate representation of the text data. (5) The detailed pre-processing workflow can be found in the supplemental code book, which is available online at https://doi.org/10.1016/j.cmpbup.2023.100104.

### 4.1.3 Text Feature Analysis – Word Frequency Score

To gain a deeper understanding of the word distribution between positive (fracture) and negative (non-fracture) reports, we computed the average Word Frequency Score (WFS) for each keyword. WFS is a metric that represents the normalized frequency of a word in a set of documents, which is calculated by dividing the total number of times a word appears in the reports by the total number of reports.

To compute the WFS, we first separated the keywords into two groups: those that appeared in positive (fracture) reports and those that appeared in negative (non-fracture) reports. For each group, we calculated the WFS for each keyword by taking the sum of the keyword frequencies in the group and dividing it by the number of reports in that group.

Next, we compared the WFS values of the keywords in the positive and negative groups to identify the words with the most significant frequency differences between the two groups. This analysis provided insights into the distinctive language patterns and terms used in the clinical reports for fracture and non-fracture cases, which could help inform the development of more accurate and interpretable machine learning models for classifying temporal bone fractures based on CT text reports.

By analyzing the WFS and identifying the keywords with the most substantial frequency differences between positive and negative sets, we gained a better understanding of the linguistic features that distinguish fracture and non-fracture cases. This information can be valuable for guiding the development of more effective and interpretable machine learning models as well as assisting physicians in recognizing the critical language cues associated with different diagnostic outcomes.

### 4.1.4 Machine Learning Model Development

To convert text reports into matrix formats suitable for machine learning models, we employed two popular text representation techniques: the bag-of-words (BOW) model and term frequency-inverse document frequency (TF-IDF) [37, 40]. Both BOW and TF-IDF methods convert each document into a fixed-length vector, enabling machine learning models to process the text in a vectorized form. In this representation, each unique word is considered a feature, so they can be recognized as an appropriate set in machine learning models for training.

Word2vec is another widely used technique for learning word associations in large text corpora. However, we believe that BOW and TF-IDF are better suited for text classification tasks in our study. The bag-of-words model helps determine a document's topic based on the types of words it contains, while the TF-IDF metric measures word relevance in the context of the entire corpus. As fracture descriptions tend to be distinct, TF-IDF can capture this characteristic: certain words frequently appear in fracture reports but rarely in non-fracture reports. In contrast, Word2vec is more suitable for discovering sub-topics or capturing semantic similarities between words, which is not the primary focus of our study. Based on these considerations, we concluded that the BOW and TF-IDF models are the most appropriate methods for the analysis.

After representing the text data using BOW and TF-IDF techniques, we trained various machine learning algorithms, such as XGBoost, Support Vector Machine, Logistic Regression, and Random Forest, to classify temporal bone fractures based on the CT text reports. By comparing the performance of these algorithms, we aimed to identify the most suitable and accurate model for fracture classification while also ensuring that the chosen model is interpretable and reliable for use in clinical settings.

### 4.1.5 Interpretation of Machine Learning Models: LIME

LIME, or Local Interpretable Model-Agnostic Explanations, is a technique developed to provide insights into the predictions made by complex machine learning models. It does so by generating simple, interpretable, and local explanations for individual predictions, allowing users to understand and trust the decisions made by the model.

In the context of classifying temporal bone fractures using CT reports, LIME is employed to identify and highlight the keywords in the text that contribute the most to the model's prediction. The method works by first training a machine learning classifier to distinguish between bone fracture (positive) and non-bone fracture (negative) cases based on the distribution of words in clinical texts. Then, LIME generates explanations by creating a simpler, interpretable white-box model that approximates the original black-box model locally around the specific prediction.

To evaluate the accuracy of LIME's explanations, two metrics are utilized: accuracy score and Kullback-Leibler (KL) divergence. The accuracy score measures the similarity between the generated sample and the original documents, with a higher score indicating a better match. The KL divergence quantifies the difference between the interpretable white-box model and the original black-box model in terms of their classification results. A lower KL divergence score signifies that the two models are more closely aligned, with a score of zero indicating a perfect match.

By using LIME, practitioners can gain a better understanding of the factors driving the predictions made by machine learning models for temporal bone fracture classification, ultimately increasing trust in and transparency of these models in clinical settings.

## 4.2 Results

### 4.2.1 Word Frequency Score and Clinical Text Summary

We began by summarizing the clinical documents. Among the 164 selected text documents, 45 were diagnosed with a bone fracture, and 119 were diagnosed without a fracture. Notably, the positive CT reports had an average length of 299.8 words (standard deviation [SD] = 124.3), which was significantly shorter than normal CT reports (average = 480.6, SD = 235.9).

In normal reports, the top five most common words were 'normal' (total frequency = 487), 'right' (393), 'canal' (356), 'CT' (347), and 'left' (334). In contrast, the top five most common words in fracture reports were 'left' (432), 'fracture' (394), 'right' (381), 'bone' (337), and 'temporal' (312).

Figure 6 illustrates the word lists that exhibit the largest word frequency score gaps between the two categories, highlighting the words favored in normal reports and those favored in fracture reports. This analysis helps identify the specific language patterns and terms that are most associated with each category, thereby shedding light on the features that the machine learning models might be leveraging to make accurate predictions.

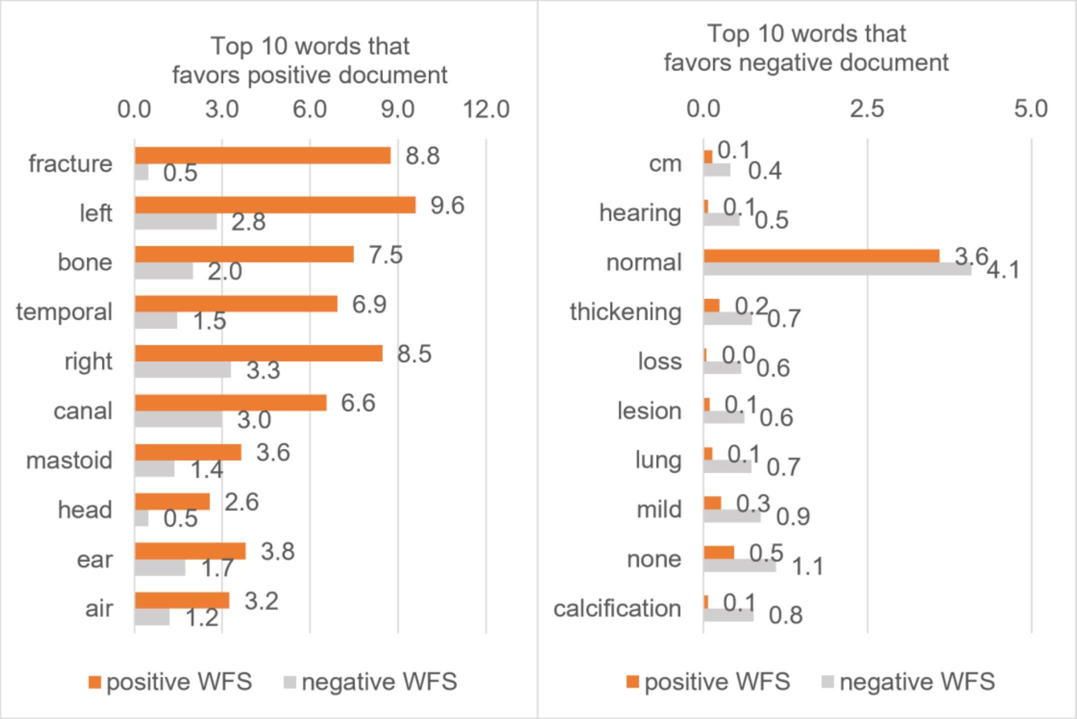


Fig. 6. Comparison of gaps between fracture and non-fracture reports. The red bar stands for the frequency of fracture reports, and the blue bar stands for non-fracture reports. The left-side chart shows the top ten words that appear more often in fracture sets, whereas the right-side chart shows the top ten words that appear more often in non-fracture sets.

### 4.2.2 Classification Models' Parameters and Performances

Figure 7 illustrates the relationship between classification model performance and the number of keywords used in the models. Each sub-figure employs either the random forest, SVM, or logistic regression algorithms. Figure 7 demonstrates a positive correlation between the number of keywords and classification performance.

As the number of keywords increases in the Random Forest model, precision and accuracy remain high, but recall begins to decline. In contrast, the SVM and logistic regression models do not exhibit a decreasing trend. Taking into account the relationship between the number of selected keywords and performance, we ultimately incorporated 500 keywords into the feature set.

Appendix G presents a table of the exact machine learning performance metrics for the different models. These results provide insight into the strengths and weaknesses of each approach, allowing for a more nuanced understanding of how the chosen features impact the classification's performance.

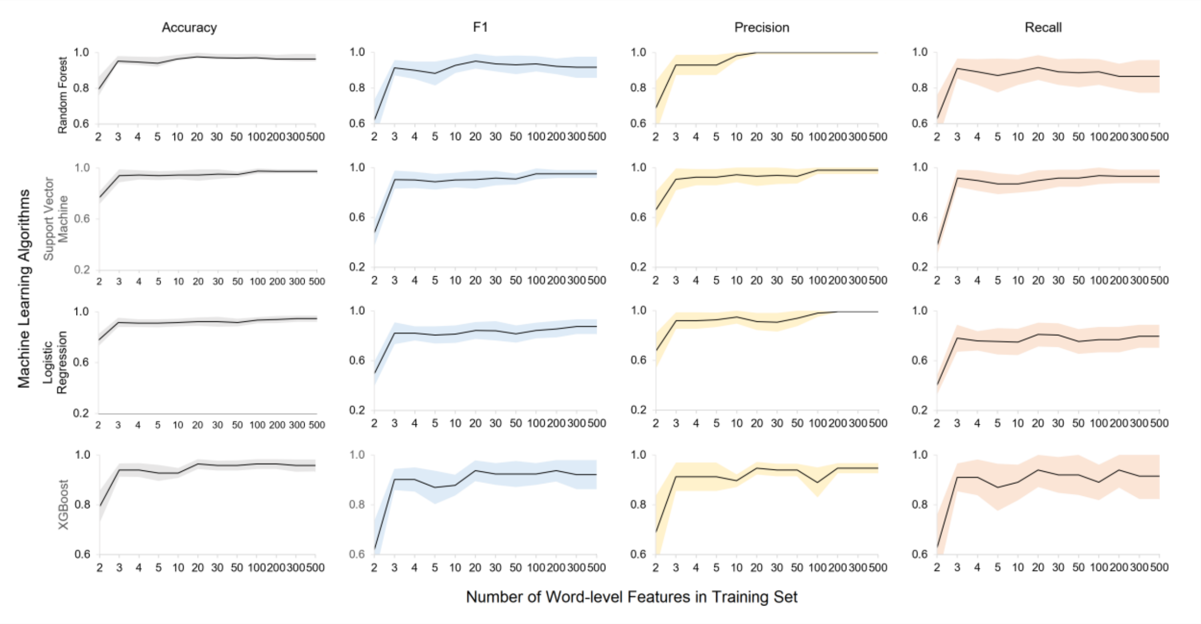


**Fig. 7.** relationships between classification model's performance, number of selected features, and evaluation performances for random forest, support vector machine, and logistic regression model.

### 4.2.3 Feature Importance

In the context of text classification, feature importance refers to a score that represents the significance of each feature, such as words or phrases, in the classification model. A higher score indicates that the specific feature has a greater impact on the model's ability to predict a specific variable, such as whether a patient has a bone fracture or not.

Understanding feature importance is crucial in text classification, as it helps identify the most relevant words or phrases that contribute to the model's decision-making process. This information can be invaluable for refining models, identifying potential biases, and providing insights into the underlying patterns present in the data.

To assess feature importance in our study, we used the mean decrease in impurity (Gini) importance score, which is a widely used measurement for tree-based models. The mean decrease in impurity is calculated as the probability of mislabeling an element, assuming that the element is randomly labeled according to the distribution of all classes in the set. For regression tasks, the analogous metric to the Gini index would be the residual sum of squares.

Figure 8 presents the top 20 most important words that contribute to the classification results, based on the mean decrease in impurity (Gini) importance score and the random forest algorithm. By identifying these key words, we can gain a better understanding of the features that the model relies on to make accurate predictions and improve the model's overall performance.

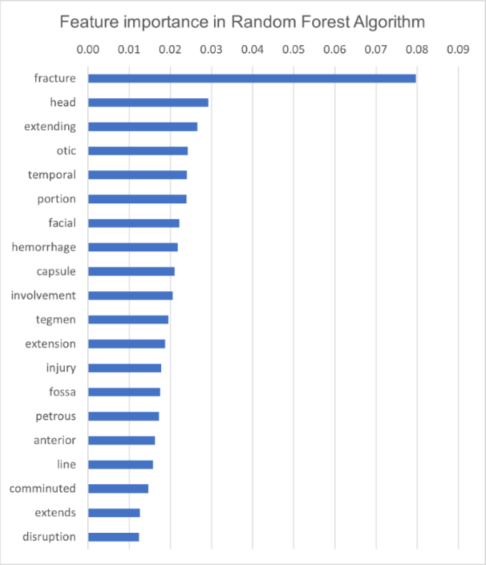
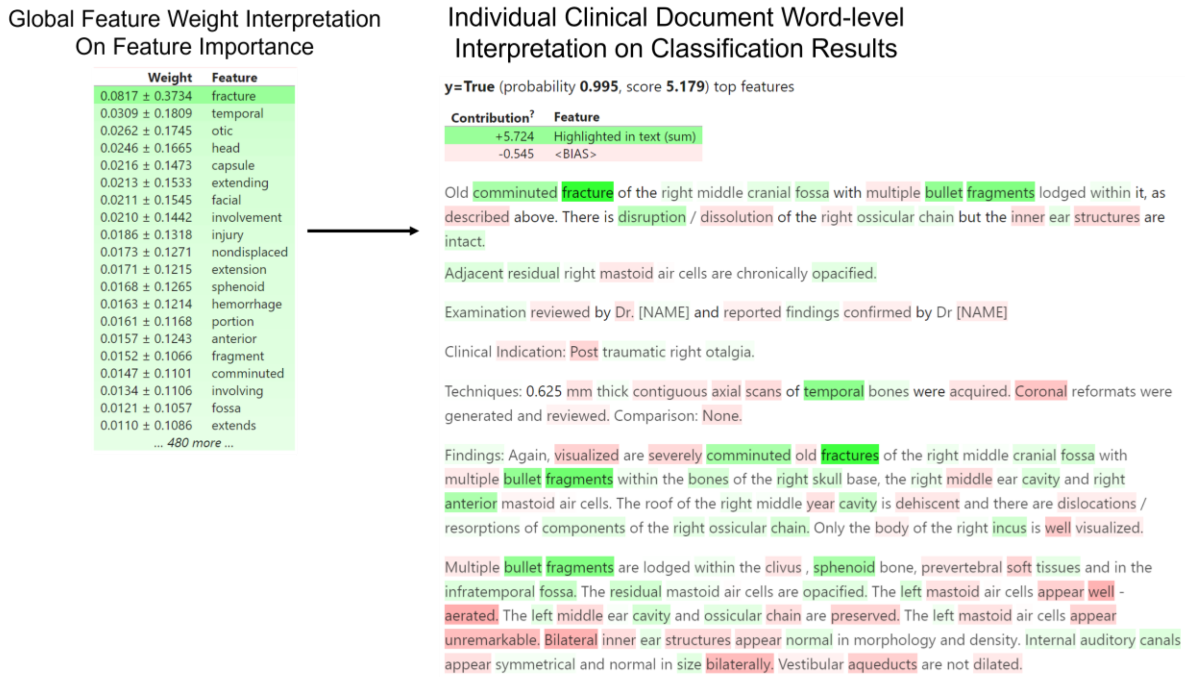


**Fig. 8.** The top 20 most important words that contribute to the classification results, based on the mean decrease in impurity (gini) importance score and random forest algorithm.

### 4.2.4 Interpretation of Machine Learning Models

Fig. 9 demonstrates a LIME Text Explainer visualization for a bone clinical text case, which highlights text features that positively or negatively influence the classification. In this example, the Random Forest algorithm was employed to generate explainable results. The visualization emphasizes the essential keywords that contribute to the final classification outcome. It is worth noting that a similar visualization was created using support vector machine algorithms, which yielded a slightly different set of keywords in the keyword feature sets.

The Random Forest classifier predicts a fracture result with 99.5% certainty and a z-score of 5.179. Words displayed in green are considered to have contributed to the model's positive classification result. In this instance, the words 'comminuted,' 'fracture,' 'lodged,' 'fossa,' 'disruption,' 'ossicular,' and 'temporal' were ranked as the most predictive words for the positive classification outcome.



Global Feature Weight Interpretation
On Feature Importance

| Weight | Feature |
|---|---|
| 0.0817 ± 0.3734 | fracture |
| 0.0309 ± 0.1809 | temporal |
| 0.0262 ± 0.1745 | otic |
| 0.0246 ± 0.1665 | head |
| 0.0216 ± 0.1473 | capsule |
| 0.0213 ± 0.1533 | extending |
| 0.0211 ± 0.1545 | facial |
| 0.0210 ± 0.1442 | involvement |
| 0.0186 ± 0.1318 | injury |
| 0.0173 ± 0.1271 | nondisplaced |
| 0.0171 ± 0.1215 | extension |
| 0.0168 ± 0.1265 | sphenoid |
| 0.0163 ± 0.1214 | hemorrhage |
| 0.0161 ± 0.1168 | portion |
| 0.0157 ± 0.1243 | anterior |
| 0.0152 ± 0.1066 | fragment |
| 0.0147 ± 0.1101 | comminuted |
| 0.0134 ± 0.1106 | involving |
| 0.0121 ± 0.1057 | fossa |
| 0.0110 ± 0.1086 | extends |
| *... 480 more ...* | |

Individual Clinical Document Word-level
Interpretation on Classification Results

**y=True** (probability **0.995**, score **5.179**) top features

| Contribution? | Feature |
|---|---|
| +5.724 | Highlighted in text (sum) |
| -0.545 | <BIAS> |

Old comminuted fracture of the right middle cranial fossa with multiple bullet fragments lodged within it, as described above. There is disruption / dissolution of the right ossicular chain but the inner ear structures are intact.

Adjacent residual right mastoid air cells are chronically opacified.

Examination reviewed by Dr. [NAME] and reported findings confirmed by Dr [NAME]

Clinical Indication: Post traumatic right otalgia.

Techniques: 0.625 mm thick contiguous axial scans of temporal bones were acquired. Coronal reformats were generated and reviewed. Comparison: None.

Findings: Again, visualized are severely comminuted old fractures of the right middle cranial fossa with multiple bullet fragments within the bones of the right skull base, the right middle ear cavity and right anterior mastoid air cells. The roof of the right middle year cavity is dehiscent and there are dislocations / resorptions of components of the right ossicular chain. Only the body of the right incus is well visualized.

Multiple bullet fragments are lodged within the clivus , sphenoid bone, prevertebral soft tissues and in the infratemporal fossa. The residual mastoid air cells are opacified. The left mastoid air cells appear well - aerated. The left middle ear cavity and ossicular chain are preserved. The left mastoid air cells appear unremarkable. Bilateral inner ear structures appear normal in morphology and density. Internal auditory canals appear symmetrical and normal in size bilaterally. Vestibular aqueducts are not dilated.

**Fig. 9.** How LIME evaluates the importance of each word features and use the weight of features to visualize the word-level contribution for each document to calculate classification results based on the random forest model.

We provided a visualization of the text explainer, where each word is assigned a contribution score indicating its impact on the positive or negative classification. The evaluation of the text explainer is done at the individual text level. The feature list highlights the most crucial words according to the random forest model, which is aggregated from multiple decision trees. The inclusion of each word in the list is based on its role as a deciding factor within a decision tree. Words with higher weight are frequently used as key factors in classification results. Both the feature list and the text explainer's assessment indicate that LIME is effective in identifying keywords for classification.

The LIME interpretation framework's reliability was evaluated using the accuracy score and the Kullback-Leibler divergence score, comparing it to the machine learning model. The accuracy score between our Random Forest model and the explainable model was found to be 0.867, meaning that 86.7% of the reports will generate the same prediction result between the two classifiers. This demonstrates a strong alignment between the two models in terms of their predictions.

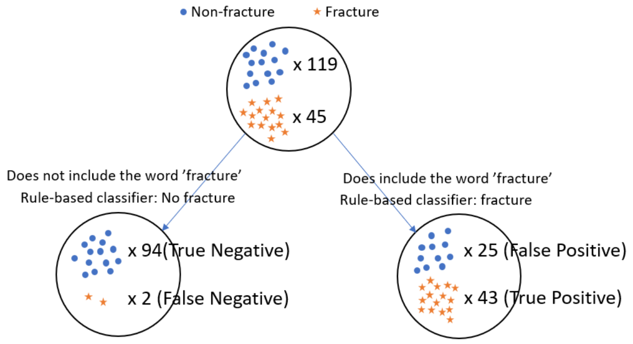
The Kullback-Leibler divergence score measures how well the probabilities are approximated between the two models for all target classes. For the SVM model, the Kullback-Leibler divergence value is 0.985, indicating that there is a 98.5% probability that both models will classify the same report into the same categories. These evaluations suggest that the Text Explainer model is highly reliable and can accurately predict the behavior of support vector machine models in CT classification tasks. Similarly, other algorithms also exhibit high reliability scores for model interpretation, further validating the trustworthiness of the LIME framework in this context.

### 4.2.5 A Comparison with a Simple Rule-Based Model

Fig. 6 suggests that the high precision probably comes from the fact that fracture reports typically contain words like 'fracture' or 'temporal.' Also, words like 'lung' or 'calcification' imply a non-fracture case because the CT images are of the lungs or the heart. It shows a sign that perhaps a simple rule-based model for these specific words may suffice for the classification task. Therefore, we included a simple rule-based model to compare with the existing models.

We use the word "fracture" to build a simple one-rule classifier. We wanted to keep the rule simple because it is common to conduct keyword searches and quickly determine the classifications. This would be a good baseline to reflect the actual scenarios for bone fracture classification.

In this case, the rule-based classifier would classify documents with "fracture" as positive and documents without "fracture" as negative. This would serve as a baseline model. In supplemental files, we included the rule-based classifier and used the "if" condition to construct the classifier on our documents. We then count the true positive, false positive, true negative, and false negative cases. We measure F1, precision, and recall in Fig. 10:


● Non-fracture
★ Fracture
x 119
x 45
Does not include the word 'fracture'
Rule-based classifier: No fracture
Does include the word 'fracture'
Rule-based classifier: fracture
x 94(True Negative)
x 2 (False Negative)
x 25 (False Positive)
x 43 (True Positive)

When we adopt "fracture" as the only rule of the rule-based classifier:

Precision = TP/(TP+FP) = 43/(43+25) = 0.632

Recall = TP/(TP+FN) = 43/(43+2) = 0.956

F1 = 2*Precision*recall/(precision + recall) = 0.761

**Fig. 10.** classification and performances of a rule-based model on 164 clinical reports

Also, we provided a supplemental file named "rule-based classifier" that provided case-by-case prediction results and rule-based model details.

From the rule-based model’s result, the F1-score is 0.761, which is significantly lower than our machine learning-based classifier. In TP, FP, FN, and TN cases, we see more false positive cases than false negative cases. This indicates even a simple rule-based classifier will not be likely to miss a fracture diagnosis. The precision is a bit lower than our machine learning model. To increase precision, therefore, it makes sense to build more complicated rules or to adopt the most frequent words as machine learning features.

As we investigated in related work, the development of rule-based classifiers occurred mostly in the 1990s. It must be admitted that a simple rule-based classifier cannot handle complex clinical text classification tasks. If we apply multiple rules, the performance may improve, but interpretation becomes a problem again. The rule-based classifier may not adapt to the ever-changing word usage in medical documents. Machine learning models, however, can overcome this problem. Therefore, building more complicated rules is no longer the focus of our study. We may not use a 20-year-old rule-based classifier as a baseline. Instead, starting with the machine learning model would be a better choice. Based on the rule-based classifier’s performance, the ML model’s development convenience, and comparisons, we decided not to include the rule-based classifier as a baseline model.

## 4.3 Discussion

### 4.3.1 Study Significance

The significance of this study lies in the effective utilization of clinical text reports, which are often an underexplored resource in electronic health records. By developing an interpretable model, medical practitioners can make more informed decisions and have increased trust in the model's predictions. This trust can help speed up the adoption of computer-based diagnoses in clinical practice.

In this study, we successfully trained machine learning classifiers to identify bone fracture cases using untemplated CT narratives from electronic health records, achieving better accuracy than similar studies that relied on crowdsourcing methods [80]. Additionally, we presented word-level contributions to the predictions using the LIME framework, which can help clinicians validate the model's validity and aid in their decision-making process.

By leveraging the LIME-based model to explain word-level contributions in clinical text, physicians can better understand how specific words, such as "fragment" and "hemorrhage," contribute to bone fracture predictions. This understanding aligns with physicians' domain knowledge, enabling them to accept the model's predictions more readily when they can see how the algorithm interprets documents and makes predictions. Providing such interpretations can increase clinical acceptance of automated systems.

Ultimately, the interpretation of a model that represents the algorithm helps foster transparency and trust among physicians. This transparency is crucial for facilitating the transition from manual to automated processes in clinical settings, ultimately improving diagnostic efficiency and patient care.

### 4.3.2 Summary of Text Feature Analysis

We discovered that text features contribute to the prediction of results. The word usage demonstrates a significant difference between positive and negative reports. For example, the word ’fracture’ was identified as the most important feature in our evaluation; it appears frequently in the positive set (frequency = 394) but infrequently in the negative data set (55). "Head" (frequency = 116 in the positive set, 57 in the negative set), "temporal" (312 in the positive set, 173 in the negative set), and "hemorrhage" are other examples (87 in the positive set, 17 in the negative set). All these words demonstrate the frequency difference between positive and negative sets. We conclude that the WFS gap between fracture and non-fracture reports can measure how the words are used differently.

In Fig. 2, the WFS result indicates that "fracture," "left," "bone," "temporal," and "right" are the top five words that appear more frequently in positive sets than in negative sets. The top five words that appear more frequently in negative sets than in positive sets are "calcification," "none," "mild," "lung," and "lesion." The difference in WFS between fracture and non-fracture reports is a predictor of classification. Physicians often draft reports with highly specialized medical terms. These medical terms often serve as reliable predictors. According to our results, we suggest investigating if non-experts can easily interpret the medical terms.

In other related studies, similar clinical text features were also examined. The goal is to investigate radiologists’ preferences for specific words in clinical documents. A previous study (, for example, used naive Bayes-based predictive machine learning models. Language patterns in clinical documents are typically consistent across specialties. The study discovered that otolaryngologists use distinct language patterns in vestibular notes that are highly conserved.

These patterns are highly predictive of specific vestibular diagnoses. Using a medically specialized corpus makes it easy for doctors to understand how language patterns work.

We believe that similar language patterns exist in other medical departments. According to the WFS gaps in Fig. 2, we classified words as fracture-favored or non-fracture-favored. The classification yields two distinct word lists. Documents with a bone fracture prefer one list of words, while documents without a fracture prefer the other. As a result, our WFS analysis identified text patterns associated with classification results. Incorporating electronic health records into decision-making models has been used to treat a variety of diagnoses and conditions, including heart failure symptoms [163], vestibular diagnoses [164], and gastrointestinal diagnoses [165].

The LIME results indicate that specialized medical words significantly contribute to the classification process. Consequently, interpretable AI has the potential to assist in explaining more complex diagnostic conditions. The pipeline developed in this study can also be used to interpret other clinical texts, classify diagnoses, and provide explanations that are easy to understand. By leveraging such interpretable AI models, medical professionals can make more informed decisions and ultimately improve patient care.

### 4.3.3 Summary of Classification Performances

In this study, it has been shown that a model utilizing 500 major topic words and stratified 10-fold cross-validation can achieve an average accuracy of 0.95 on Random Forest classifiers. This performance is competitive compared to other human-labeled studies, such as the crowdsourcing method that achieved 0.799 accuracy [80]. Although the model exhibits high precision, its recall is lower, indicating a tendency to predict more positive outcomes than negative ones. Consequently, the model may help reduce false-negative cases, potentially avoiding serious errors in clinical practice.

A positive correlation was observed between the number of keywords used in the feature set and model performance across all four algorithms. Increasing the number of features can improve prediction performance, particularly when the number of keywords is below ten. However, performance plateaus once the number of keywords surpasses two hundred. This plateau is likely due to the limited impact of less-important keywords with lower WFS, which minimally contribute to classification.

In summary, all four algorithms—random forest, support vector machine, logistic regression, and XGBoost—can perform classifications with high accuracy. The performance can be enhanced by increasing the number of features in the model, although the impact diminishes beyond a certain point. By utilizing these algorithms and adjusting the number of features, machine learning models effectively classify clinical texts and improve decision-making in medical practice.

### 4.3.4 Summary of Interpretation Methods

LIME has indeed been widely used in various research studies, as it provides a visualized and easily interpretable explanation for predictions made by machine learning models. This has made it a popular choice for interpreting clinical decision models across a range of medical conditions and diagnostics. Liyan and Mao et al. [78] used LIME to investigate the level of contribution of features in new instances for predicting central precocious puberty in girls. Ghafouri-Fard et al. [166] used the same approach for diagnosing autism spectrum disorder. Palatnik de Sousa et al. used LIME to classify the metastases of lymph nodes [167]. Other interpretations of target conditions using LIME include acute kidney injury [137], chest injury [168], electrocardiogram-aided cardiovascular diseases [149], radiology reports [116], [122], and so on. Overall, LIME can provide visualized results for various diagnoses to help clinicians evaluate the reliability of clinical decision models. There are two main approaches for implementing classification model interpretation:

Intrinsic transparency: This approach uses inherently interpretable models, such as decision trees, which are easy to understand and explain by design. The benefit of intrinsically transparent models is that they provide an immediate understanding of the decision-making process. We list pros and cons for each approach. For intrinsic transparency, it allows a direct understanding of the decision-making process and does not need additional explanation methods; for cons, intrinsic transparency may have lower performance compared to more complex models and possess limited flexibility in modeling complex relationships.

A post-hoc interpretation, which includes methods like LIME, provides explanations for predictions made by more complex models, such as deep learning or ensemble methods. These methods are applied after the model has been trained and used for prediction, hence the term "post-hoc." The pros include their applicability to various complex models, even those that are not intrinsically transparent. They not only provide detailed insights into the features that contribute to predictions but also enable the use of high-performing models while still providing interpretability. However, the cons include the fact that post-hoc interpretations may not always be perfect or completely accurate. They require additional computation and implementation efforts. Besides, while they are much easier to understand than the algorithms of machine learning models, they may still be challenging for non-experts to understand in some cases.

When choosing ML-based explanation methods for future applications, it is essential to weigh the pros and cons of each implementation approach. Ultimately, the choice will depend on the specific use case, the desired level of interpretability, and the trade-offs between model performance and interpretability.

### 4.3.5 Interpretability of Machine Learning Models

The interpretability of machine learning models plays a crucial role in their adoption in real-world applications, especially in the medical domain. Some machine models are inherently transparent and interpretable due to their simple mechanisms, which makes their output directly interpretable. For instance, linear regression and logistic regression are more transparent by nature, and clinicians have extensive experience interpreting their coefficients, effect sizes, and p-values. Our previous studies explored the social determinants of tertiary rhinology care utilization using linear regression techniques, demonstrating that no AI-based knowledge is necessary for understanding such models.

In contrast, decision trees are slightly more complex but still transparent models [125]. In machine learning, decision trees are used to define a preferred sequence of attributes for investigation, which narrows down to a specific outcome or state. This process, known as decision tree learning, relies on selecting attributes with high mutual information. A higher information gain is applied to the split for each node, and as we mentioned earlier, we can calculate the information gain for specific words. Decision trees used for text classification consist of internal tree nodes labeled by terms, branches departing from them labeled by tests on the weight, and leaf nodes representing corresponding class labels. Decision trees classify documents based on predetermined rules by traversing the query structure from root to leaf, which is the ultimate goal of classification.

Fig. 10 presents a simplified decision tree illustrating the process of making a diagnosis based on the frequency of specific words in CT reports. This decision tree is generated from an example document, with each square indicating a criterion when a document is used as input. The number in each square represents the frequency of words occurring in the document, and each conditional criterion is followed by a percentage number indicating the proportion of documents that satisfy that condition. Ultimately, each document is classified into one of the categories, resulting in a "positive" or "negative" outcome. Decision rules are learned using machine learning techniques and information gain, which requires some statistical knowledge to build the decision tree. Fortunately, decision rules and sequences are directly interpretable by clinicians, especially if the tree is small.

A significant difference exists between decision tree and LIME methods in terms of interpretation complexity. While decision trees require clinicians to analyze the reasonableness of an entire tree's sequence, LIME methods only necessitate determining if featured words are associated with target outcomes. However, decision trees can become too large and unwieldy to interpret, or they may not generate a meaningful representation. Consequently, LIME methods offer a significantly easier approach to interpreting models, enabling greater adoption and understanding of machine learning models in clinical practice.

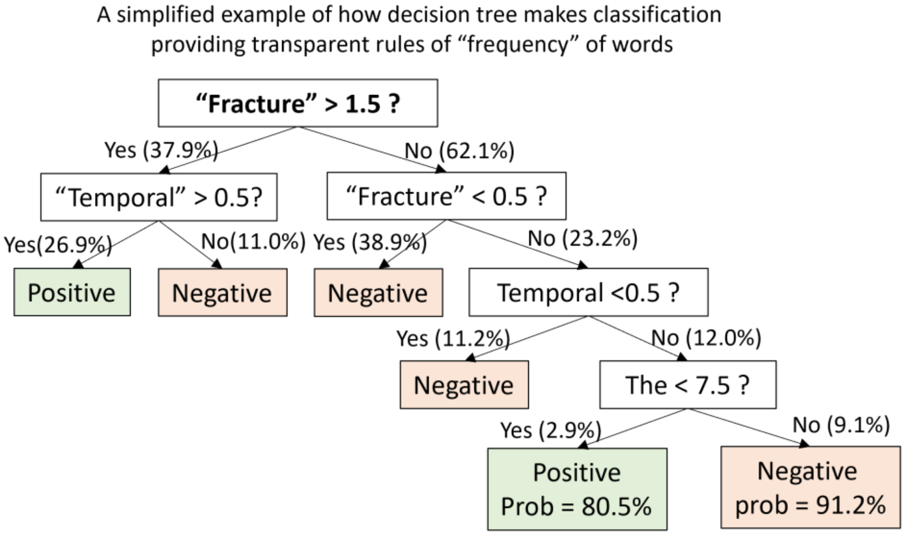


**Fig. 11.** A visualization of how a transparent decision tree model determines the classification results from a sample CT text report. The percentage number shows the proportion of reports falling into each category. The frequency of a specific word determines the model's classification result.

### 4.4 Limitations and Future Work

We acknowledge some limitations. The first limitation is the limited variety and quantity of temporal bone CT reports, with only 164 documents available. All reports were limited to a single health care system in Wisconsin, which may introduce potential bias. A larger set of clinical reports may also lead to unbiased model construction and more accurate classification performance. The age range was limited to 60–65 years old, which may affect the generalizability of the conclusion. Our future work will include two aspects: First, we used labeled data in this preliminary study. While unlabeled data cannot be used for classification, it has the potential for unsupervised learning. We believe that by building an appropriate unsupervised model, it is possible to cluster CT reports into two categories based on text reports. Second, building a medically specialized text interpreter would highlight only medical words and achieve a more transparent interpretation. For example, by adopting the SNOMED-CT standard [169], we can create a medical text interpreter. By using only medical terms, the model could narrow down the choices of words. The word-level optimization may achieve more accurate prediction and interpretation.

### 4.5 Conclusion

Machine learning has shown significant success in classifying patients' diagnostic outcomes using free-text clinical notes. However, a major challenge to adopting these models in clinical practice is the interpretation of their algorithms. This study aimed to address this challenge by presenting four classifiers for the classification of fracture cases from untemplated temporal bone CT reports.

The study's main contributions are threefold: (1) A text analysis revealed significant differences in word usage between fracture and non-fracture report sets, highlighting the importance of understanding the language patterns in clinical documents; (2) Random Forest-based classification achieved the highest accuracy among the four algorithms used, demonstrating the potential of machine learning in the medical field; (3) A LIME-based approach provided interpretable explanations by visualizing the contribution of words in bone fracture classification, enhancing trust and transparency in computerized models.

Understanding the decision-making mechanisms behind these models is crucial for promoting their use in clinical contexts. The proposed approach can support clinical decision-making by providing simple visualizations to physicians, helping them validate the model's validity and reliability. This increased trust in computerized models can facilitate the adoption of these tools in daily practice, serving as a complementary aid for CT report classification and assisting clinicians in making more informed decisions.

Overall, this study laid the groundwork for the development and validation of reliable explanations in machine learning-based clinical models. By undertaking further research in realistic scenarios and real-world settings, the model has the potential to be integrated into contemporary medical decision-making environments for clinical practitioners. Future work will focus on leveraging unsupervised learning to cluster CT reports, building medically specialized text interpreters using standardized medical terms, and addressing the limitations identified in the study, such as the limited quantity and variety of CT reports. These advancements will contribute to more accurate predictions and interpretations, further bridging the gap between machine learning models and their application in clinical practice.

# 5. Conclusion

### 5.1 Summary of Two Studies

Falls among older adults have emerged as a significant public health concern in the United States, with one in four older adults reporting a fall each year. The consequences of falls can be severe, leading to approximately 36 million falls and over 32,000 deaths annually. As the U.S. population continues to age, the risk of falls and associated problems is likely to increase. Our research efforts aimed to address this issue from two distinct healthcare perspectives: developing new AI-based prediction algorithms to enhance the quality of care and examining health equity problems in an economically diverse society to ensure high-quality care for patients at risk of falls.

In the first study, we investigated disparities in fall patients' access to healthcare services. We discovered that unequal access to care may disproportionately affect certain patient subsets, particularly those with lower socioeconomic status. This highlights the need for further investigation to identify the underlying causes of such disparities and develop targeted interventions. We posit that effective preventive strategies will require collaboration among medical facilities, governments, and local clinical service providers to improve older adults' access to quality care, thereby mitigating the risks and consequences of falls.

In the second study, we explored the interpretability of machine learning models in the context of medical documents, recognizing that such models are often opaque and challenging for clinical providers to understand. Addressing the need for more accessible interpretations of clinical texts, we employed two key methodologies in this study: (1) calculating average word frequency scores for essential keywords and (2) utilizing Local Interpretable Model-agnostic Explanations (LIME) to visualize the contribution weight of keywords to bone fracture diagnoses. Our findings suggest that interpretable text explainers can enhance physicians' comprehension of machine learning predictions, fostering greater trust in computerized models and facilitating their use in clinical decision-making processes. Understanding the decision-making system's mechanism is critical for promoting its use in clinical settings. By providing physicians with simple visualizations, our model can help them make decisions. This interpretation increases confidence in computerized models. Our proposed method could be used as a computer-assisted tool in CT report classification. It could be used as an adjunct tool to assist clinicians in making decisions in their daily practice. Overall, this study laid the groundwork for the development and validation of credible explanations. Our model has the potential to be integrated into contemporary clinical decision-making environments for clinical practitioners.

In summary, our research contributes to a broader understanding of fall risks among older adults by highlighting healthcare disparities and offering actionable insights for improving access to care. Furthermore, we demonstrate the potential of interpretable machine learning models to support clinical decision-making, emphasizing the need for clear, comprehensible, and trustworthy AI-based tools in medical settings. By integrating these findings, we hope to create a foundation for future research and interventions aimed at reducing the incidence and impact of falls in the aging population, ultimately leading to better health outcomes and improved quality of life.

## 5.2 Funding

The two studies described are supported by the National Center for Advancing Translational Sciences, National Institutes of Health, Award Number UL1TR001436. The content is solely the responsibility of the author(s) and does not necessarily represent the official views of the NIH. This project is also funded by the Research and Education Program Fund, a component of the Advancing a Healthier Wisconsin Endowment at the Medical College of Wisconsin.

## 5.3 Ethics Approval

In the first study, the Medical College of Wisconsin and Froedtert Hospital Institutional Review Board (PRO00036649) approved the OTO Clinomic platform, the interrogation of the electronic medical record, and the retrospective chart review involved in this study. The second study is an observational study. The University of Wisconsin—Milwaukee Institutional Research Ethics Committee has confirmed that no ethical approval is required.

## 5.4 Declaration of Competing Interest

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

# 6. References

[1] A. F. Ambrose, G. Paul, and J. M. Hausdorff, “Risk factors for falls among older adults: a review of the literature,” *Maturitas*, vol. 75, no. 1, pp. 51–61, 2013.

[2] G. Bergen, M. R. Stevens, and E. R. Burns, “Falls and fall injuries among adults aged≥ 65 years—United States, 2014,” *Morbidity and Mortality Weekly Report*, vol. 65, no. 37, pp. 993–998, 2016.

[3] S. K. Inouye, C. J. Brown, M. E. Tinetti, and others, “Medicare nonpayment, hospital falls, and unintended consequences,” *New England Journal of Medicine*, vol. 360, no. 23, p. 2390, 2009.

[4] B. H. Alexander, F. P. Rivara, and M. E. Wolf, “The cost and frequency of hospitalization for fall-related injuries in older adults.,” *Am J Public Health*, vol. 82, no. 7, pp. 1020–1023, 1992.

[5] B. Moreland and R. Lee, “Emergency department visits and hospitalizations for selected nonfatal injuries among adults aged≥ 65 years—United States, 2018,” *Morbidity and Mortality Weekly Report*, vol. 70, no. 18, pp. 661–666, 2021.

[6] A. Subcommittee, I. A. W. Group, and others, “Advanced trauma life support (ATLS®): the ninth edition,” *J Trauma Acute Care Surg*, vol. 74, no. 5, pp. 1363–1366, 2013.

[7] M. C. Nevitt, S. R. Cummings, and E. S. Hudes, “Risk factors for injurious falls: a prospective study,” *J Gerontol*, vol. 46, no. 5, pp. M164–M170, 1991.

[8] D. Waltzman, J. Haarbauer-Krupa, and L. S. Womack, “Traumatic brain injury in older adults—a public health perspective,” *JAMA Neurol*, vol. 79, no. 5, pp. 437–438, 2022.

[9] R. W. Sattin, “Falls among older persons: a public health perspective,” *Annu Rev Public Health*, vol. 13, no. 1, pp. 489–508, 1992.

[10] F. Li and P. Harmer, “Prevalence of falls, physical performance, and dual-task cost while walking in older adults at high risk of falling with and without cognitive impairment,” *Clin Interv Aging*, pp. 945–952, 2020.

[11] T. Karpusenko *et al.*, “Factors associated with unrecovered falls among older adults,” *Geriatr Nurs (Minneap)*, vol. 51, pp. 323–329, 2023.

[12] M. C. Hornbrook, V. J. Stevens, D. J. Wingfield, J. F. Hollis, M. R. Greenlick, and M. G. Ory, “Preventing falls among community-dwelling older persons: results from a randomized trial,” *Gerontologist*, vol. 34, no. 1, pp. 16–23, 1994.

[13] A. C. Scheffer, M. J. Schuurmans, N. van Dijk, T. van der Hooft, and S. E. de Rooij, “Fear of falling: measurement strategy, prevalence, risk factors and consequences among older persons,” *Age Ageing*, vol. 37, no. 1, pp. 19–24, Apr. 2008, doi: 10.1093/ageing/afm169.

[14] E. W. Gregg, M. A. Pereira, and C. J. Caspersen, “Physical activity, falls, and fractures among older adults: a review of the epidemiologic evidence,” *J Am Geriatr Soc*, vol. 48, no. 8, pp. 883–893, 2000.

[15] N. R. Nicholson, “A review of social isolation: an important but underassessed condition in older adults,” *J Prim Prev*, vol. 33, pp. 137–152, 2012.

[16] J. Vespa, D. M. Armstrong, L. Medina, and others, *Demographic turning points for the United States: Population projections for 2020 to 2060*. US Department of Commerce, Economics and Statistics Administration, US~…, 2018.

[17] R. Camicioli and S. R. Majumdar, “Relationship between mild cognitive impairment and falls in older people with and without Parkinson’s disease: 1-Year Prospective Cohort Study,” *Gait Posture*, vol. 32, no. 1, pp. 87–91, 2010.

[18] S. D. Berry and R. R. Miller, “Falls: epidemiology, pathophysiology, and relationship to fracture,” *Curr Osteoporos Rep*, vol. 6, no. 4, pp. 149–154, 2008.

[19] M. K. Appeadu and B. Bordoni, “Falls and fall prevention in the elderly,” in *StatPearls [Internet]*, StatPearls Publishing, 2022.

[20] T. Al-Aama, “Falls in the elderly: spectrum and prevention,” *Canadian Family Physician*, vol. 57, no. 7, pp. 771–776, 2011.

[21] C. Sherrington *et al.*, “Exercise for preventing falls in older people living in the community,” *Cochrane database of systematic reviews*, no. 1, 2019.

[22] F. Bunn, A. Dickinson, E. Barnett-Page, E. Mcinnes, and K. Horton, “A systematic review of older people’s perceptions of facilitators and barriers to participation in falls-prevention interventions,” *Ageing Soc*, vol. 28, no. 4, pp. 449–472, 2008.

[23] C. S. Florence, G. Bergen, A. Atherly, E. Burns, J. Stevens, and C. Drake, “Medical costs of fatal and nonfatal falls in older adults,” *J Am Geriatr Soc*, vol. 66, no. 4, pp. 693–698, 2018.

[24] J. A. Stevens, P. S. Corso, E. A. Finkelstein, and T. R. Miller, “The costs of fatal and non-fatal falls among older adults,” *Injury prevention*, vol. 12, no. 5, pp. 290–295, 2006.

[25] N. D. A. Boyé, E. M. M. Van Lieshout, E. F. Van Beeck, K. A. Hartholt, T. J. M. der Cammen, and P. Patka, “The impact of falls in the elderly,” *Trauma*, vol. 15, no. 1, pp. 29–35, 2013.

[26] Y. Dionyssiotis, “Analyzing the problem of falls among older people,” *Int J Gen Med*, pp. 805–813, 2012.

[27] A. Pighills and L. Clemson, *Environmental risk factors for falls*. Cambridge University Press Cambridge, 2021.

[28] J. Pynoos, B. A. Steinman, and A. Q. D. Nguyen, “Environmental assessment and modification as fall-prevention strategies for older adults,” *Clin Geriatr Med*, vol. 26, no. 4, pp. 633–644, 2010.

[29] A. F. Ambrose, G. Paul, and J. M. Hausdorff, “Risk factors for falls among older adults: A review of the literature,” *Maturitas*, vol. 75, no. 1, pp. 51–61, May 2013, doi: 10.1016/J.MATURITAS.2013.02.009.

[30] L. D. Gillespie *et al.*, “Interventions for preventing falls in older people living in the community,” *Cochrane database of systematic reviews*, no. 9, 2012.

[31] A. Gangavati *et al.*, “Hypertension, orthostatic hypotension, and the risk of falls in a community-dwelling elderly population: the maintenance of balance, independent living, intellect, and zest in the elderly of Boston study,” *J Am Geriatr Soc*, vol. 59, no. 3, pp. 383–389, 2011.

[32] F. C. Pampel, P. M. Krueger, and J. T. Denney, “Socioeconomic disparities in health behaviors,” *Annu Rev Sociol*, vol. 36, pp. 349–370, 2010.

[33] A. Steptoe and P. Zaninotto, “Lower socioeconomic status and the acceleration of aging: An outcome-wide analysis,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 26, pp. 14911–14917, 2020.

[34] M. Pinquart and S. Sörensen, “Influences of socioeconomic status, social network, and competence on subjective well-being in later life: a meta-analysis.,” *Psychol Aging*, vol. 15, no. 2, p. 187, 2000.

[35] P. Braveman, “What are health disparities and health equity? We need to be clear,” *Public health reports*, vol. 129, no. 1_suppl2, pp. 5–8, 2014.

[36] P. A. Braveman, C. Cubbin, S. Egerter, D. R. Williams, and E. Pamuk, “Socioeconomic disparities in health in the United States: what the patterns tell us,” *Am J Public Health*, vol. 100, no. S1, pp. S186–S196, 2010.

[37] H. K. Koh, S. C. Oppenheimer, S. B. Massin-Short, K. M. Emmons, A. C. Geller, and K. Viswanath, “Translating research evidence into practice to reduce health disparities: a social determinants approach,” *Am J Public Health*, vol. 100, no. S1, pp. S72–S80, 2010.

[38] L. A. Penner, N. Hagiwara, S. Eggly, S. L. Gaertner, T. L. Albrecht, and J. F. Dovidio, “Racial healthcare disparities: A social psychological analysis,” *Eur Rev Soc Psychol*, vol. 24, no. 1, pp. 70–122, 2013.

[39] W. H. Organization, W. H. Organization. Ageing, and L. C. Unit, *WHO global report on falls prevention in older age*. World Health Organization, 2008.

[40] A. D. Thierry, “Association between telomere length and neighborhood characteristics by race and region in US midlife and older adults,” *Health Place*, vol. 62, p. 102272, 2020.

[41] J. A. Gliedt, A. L. Specto, M. J. Schneider, J. Williams, and S. Young, “Disparities in chiropractic utilization by race, ethnicity and socioeconomic status: A scoping review of the literature,” *J Integr Med*, 2023.

[42] A. Schulz and M. E. Northridge, “Social determinants of health: implications for environmental health promotion,” *Health education & behavior*, vol. 31, no. 4, pp. 455–471, 2004.

[43] R. L. Berg, J. S. Cassells, and others, “Falls in older persons: risk factors and prevention,” in *The second fifty years: Promoting health and preventing disability*, National Academies Press (US), 1992.

[44] S. Khairat *et al.*, “Advancing health equity and access using telemedicine: A geospatial assessment,” *Journal of the American Medical Informatics Association*, vol. 26, no. 8–9, pp. 796–805, Apr. 2019, doi: 10.1093/jamia/ocz108.

[45] R. I. Stone, “Successful aging in community: The role of housing, services, and community integration,” *Generations*, vol. 40, no. 4, pp. 67–73, 2017.

[46] H. Luukinen, K. Koski, R. Honkanen, and S.-L. Kivelä, “Incidence of injury-causing falls among older adults by place of residence: A population-based study,” *J Am Geriatr Soc*, vol. 43, no. 8, pp. 871–876, 1995.

[47] S. D. Berry and R. R. Miller, “Falls: epidemiology, pathophysiology, and relationship to fracture,” *Curr Osteoporos Rep*, vol. 6, no. 4, pp. 149–154, 2008.

[48] J. A. Grisso *et al.*, “Risk factors for falls as a cause of hip fracture in women,” *New England journal of medicine*, vol. 324, no. 19, pp. 1326–1331, 1991.

[49] C. R. Covert and G. S. Fox, “Anaesthesia for hip surgery in the elderly,” *Canadian Journal of Anaesthesia*, vol. 36, pp. 311–319, 1989.

[50] J. Craig, A. Murray, S. Mitchell, S. Clark, L. Saunders, and L. Burleigh, “The high cost to health and social care of managing falls in older adults living in the community in Scotland,” *Scott Med J*, vol. 58, no. 4, pp. 198–203, 2013.

[51] T. Peterson, R. Ramsey, M. Fernandez, A. Hin, C. Prado, and P. Reyes, “Engaging with First Responders to Prevent Falls in Older Adults,” 2013.

[52] S. M. Bradley, “Falls in older adults,” *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, vol. 78, no. 4, pp. 590–595, 2011.

[53] P. Zaninotto, Y.-T. Huang, G. Di Gessa, J. Abell, C. Lassale, and A. Steptoe, “Polypharmacy is a risk factor for hospital admission due to a fall: evidence from the English Longitudinal Study of Ageing,” *BMC Public Health*, vol. 20, pp. 1–7, 2020.

[54] J. M. Cancio, E. Vela, S. Santaeugènia, M. Clèries, M. Inzitari, and D. Ruiz, “Long-term impact of hip fracture on the use of healthcare resources: a population-based study,” *J Am Med Dir Assoc*, vol. 20, no. 4, pp. 456–461, 2019.

[55] D. R. Levy *et al.*, “Reflections on the Documentation Burden Reduction AMIA Plenary Session through the Lens of 25x5,” *Appl Clin Inform*, no. AAM, 2022.

[56] T. H. Payne *et al.*, “Report of the AMIA EHR-2020 Task Force on the status and future direction of EHRs,” *Journal of the American Medical Informatics Association*, vol. 22, no. 5, pp. 1102–1110, 2015.

[57] L. Liu, E. Stroulia, I. Nikolaidis, A. Miguel-Cruz, and A. R. Rincon, “Smart homes and home health monitoring technologies for older adults: A systematic review,” *Int J Med Inform*, vol. 91, pp. 44–59, 2016.

[58] G. Forbes, S. Massie, and S. Craw, “Fall prediction using behavioural modelling from sensor data in smart homes,” *Artif Intell Rev*, vol. 53, no. 2, pp. 1071–1091, Mar. 2019, doi: 10.1007/S10462-019-09687-7.

[59] N. Zerrouki, F. Harrou, A. Houacine, and Y. Sun, “Fall detection using supervised machine learning algorithms: A comparative study,” *Proceedings of 2016 8th International*

*Conference on Modelling, Identification and Control, ICMIC 2016*, pp. 665–670, Jan. 2017, doi: 10.1109/ICMIC.2016.7804195.

[60] A. O. Kansiz, M. Amac Guvensan, and H. Irem Turkmen, “Selection of time-domain features for fall detection based on supervised learning,” in *Proceedings of the World Congress on Engineering and Computer Science*, San Francisco, 2013. Accessed: Oct. 03, 2021. [Online]. Available: https://www.researchgate.net/profile/Amac-Guvensan/publication/289618532_Selection_of_Time-Domain_Features_for_Fall_Detection_Based_on_Supervised_Learning/links/5fc14bae458515b7977b841f/Selection-of-Time-Domain-Features-for-Fall-Detection-Based-on-Supervised-Learning.pdf

[61] K. Yang, C. R. Ahn, M. C. Vuran, and S. S. Aria, “Semi-supervised near-miss fall detection for ironworkers with a wearable inertial measurement unit,” *Autom Constr*, vol. 68, pp. 194–202, Aug. 2016, doi: 10.1016/J.AUTCON.2016.04.007.

[62] P. N. Ali Fahmi, V. Viet, and D. J. Choi, “Semi-supervised fall detection algorithm using fall indicators in smartphone,” *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication*, 2012, doi: 10.1145/2184751.2184890.

[63] T. Zhang, J. Wang, L. Xu, and P. Liu, “Fall Detection by Wearable Sensor and One-Class SVM Algorithm,” pp. 858–863, 2006, doi: 10.1007/978-3-540-37258-5_104.

[64] A. Núñez-Marcos, G. Azkune, and I. Arganda-Carreras, “Vision-based fall detection with convolutional neural networks,” *Wirel Commun Mob Comput*, vol. 2017, 2017, doi: 10.1155/2017/9474806.

[65] A. Cami, A. Arnold, S. Manzi, and B. Reis, “Predicting adverse drug events using pharmacological network models,” *Sci Transl Med*, vol. 3, no. 114, pp. 114ra127-114ra127, Dec. 2011, doi: 10.1126/scitranslmed.3002774.

[66] M. Clark, “Prediction of clinical risks by analysis of preclinical and clinical adverse events,” *J Biomed Inform*, vol. 54, pp. 167–173, Apr. 2015, doi: 10.1016/j.jbi.2015.02.008.

[67] N. P. Tatonetti, P. P. Ye, R. Daneshjou, and R. B. Altman, “Data-driven prediction of drug effects and interactions,” *Sci Transl Med*, vol. 4, no. 125, pp. 125ra31-125ra31, Mar. 2012, doi: 10.1126/scitranslmed.3003377.

[68] M. Saad *et al.*, “Association between COVID-19 Diagnosis and In-Hospital Mortality in Patients Hospitalized with ST-Segment Elevation Myocardial Infarction,” *JAMA - Journal of the American Medical Association*, vol. 326, no. 19, pp. 1940–1952, Nov. 2021, doi: 10.1001/jama.2021.18890.

[69] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. Feng, “Early diagnosis of Alzheimer’s disease with deep learning,” *2014 IEEE 11th International Symposium on Biomedical Imaging, ISBI 2014*, pp. 1015–1018, Jul. 2014, doi: 10.1109/ISBI.2014.6868045.

[70] M. Tariq, S. Iqbal, H. Ayesha, I. Abbas, K. T. Ahmad, and M. F. K. Niazi, “Medical image based breast cancer diagnosis: State of the art and future directions,” *Expert Systems with Applications*, vol. 167. Elsevier Ltd, Apr. 01, 2021. doi: 10.1016/j.eswa.2020.114095.

[71] L. Floridi, “What the near future of artificial intelligence could be,” *Philos Technol*, vol. 32, pp. 1–15, 2019.

[72] R. Rajagopalan, I. Litvan, and T.-P. Jung, “Fall prediction and prevention systems: recent trends, challenges, and future research directions,” *Sensors*, vol. 17, no. 11, p. 2509, 2017.

[73] E. H. Shortliffe and J. J. Cimino, *Biomedical Informatics: Computer applications in health care and biomedicine*, 4th ed. Springer, 2014.

[74] P. B. Jensen, L. J. Jensen, and S. Brunak, “Mining electronic health records: towards better research applications and clinical care,” *Nature Reviews Genetics 2012 13:6*, vol. 13, no. 6, pp. 395–405, May 2012, doi: 10.1038/nrg3208.

[75] R. Greenes, *Clinical decision support: the road to broad adoption*, Academic Press. Academic Press, 2014. Accessed: Jul. 14, 2021. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=rwrUAgAAQBAJ&oi=fnd&pg=PP1&dq=Clinical+decision+support:+the+road+to+broad+adoption&ots=mJeabVh59E&sig=b7c7SdcYt4iKrF7PSHzdPcQtupU

[76] T. A. Koleck, C. Dreisbach, P. E. Bourne, and S. Bakken, “Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review,” *Journal of the American Medical Informatics Association*, vol. 26, no. 4, pp. 364–379, Apr. 2019, doi: 10.1093/JAMIA/OCY173.

[77] Y. Shao, S. Taylor, N. Marshall, C. Morioka, and Q. Zeng-Treitler, “Clinical Text Classification with Word Embedding Features vs. Bag-of-Words Features,” *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, pp. 2874–2878, Jan. 2019, doi: 10.1109/BIGDATA.2018.8622345.

[78] Liyan *et al.*, “Development of Prediction Models Using Machine Learning Algorithms for Girls with Suspected Central Precocious Puberty: Retrospective Study,” *JMIR Med Inform 2019;7(1):e11728 https://medinform.jmir.org/2019/1/e11728*, vol. 7, no. 1, p. e11728, Feb. 2019, doi: 10.2196/11728.

[79] T. Davenport and R. Kalakota, “The potential for artificial intelligence in healthcare,” *Future Healthc J*, vol. 6, no. 2, p. 94, 2019.

[80] H. F. da Cruz *et al.*, “Using interpretability approaches to update ‘black-box’ clinical prediction models: an external validation study in nephrology,” *Artif Intell Med*, vol. 111, p. 101982, Jan. 2021, doi: 10.1016/J.ARTMED.2020.101982.

[81] P. Tamagnini, J. Krause, A. Dasgupta, and E. Bertini, “Interpreting black-box classifiers using instance-level visual explanations,” in *Proceedings of the 2nd workshop on human-in-the-loop data analytics*, 2017, pp. 1–6.

[82] K. De Groot, A. J. E. De Veer, A. M. Munster, A. L. Francke, and W. Paans, “Nursing documentation and its relationship with perceived nursing workload: a mixed-methods study among community nurses,” *BMC Nurs*, vol. 21, no. 1, pp. 1–12, 2022.

[83] L. Tong *et al.*, “Telemedicine and health disparities: Association between patient characteristics and telemedicine, in-person, telephone and message-based care during the COVID-19 pandemic,” *Ipem-translation*, vol. 3, p. 100010, 2022.

[84] J. Luo *et al.*, “Telemedicine adoption during the COVID-19 pandemic: gaps and inequalities,” *Appl Clin Inform*, vol. 12, no. 04, pp. 836–844, 2021.

[85] L. Tong, M. Khani, Q. Lu, B. Taylor, K. Osinski, and J. Luo, “Association between body-mass index, patient characteristics, and obesity-related comorbidities among COVID-19 patients: A prospective cohort study,” *Obes Res Clin Pract*, vol. 17, no. 1, pp. 47–57, 2023.

[86] D. M. Poetker, D. R. Friedland, J. A. Adams, L. Tong, K. Osinski, and J. Luo, “Socioeconomic determinants of tertiary rhinology care utilization,” *OTO Open*, vol. 5, no. 2, p. 2473974X211009830, 2021.

[87] N. K. Osafo *et al.*, “Standardization of Outcome Measures for Intratympanic Steroid Treatment for Idiopathic Sudden Sensorineural Hearing Loss,” *Otology & Neurotology*, vol. 43, no. 10, pp. 1137–1143, 2022.

[88] S. W. White *et al.*, “Analysis of socioeconomic factors in laryngology clinic utilization for treatment of dysphonia,” *Laryngoscope Investig Otolaryngol*, vol. 7, no. 1, pp. 202–209, 2022.

[89] A. Thompson-Harvey, D. R. Friedland, J. A. Adams, L. Tong, K. Osinski, and J. Luo, “The Demographics of Menière’s Disease: Selection Bias or Differential Susceptibility?,” *Otology & Neurotology*, vol. 44, no. 2, pp. e95–e102, 2023.

[90] M. A. Patel *et al.*, “Demographic differences in the treatment of unilateral vocal fold paralysis,” *Laryngoscope Investig Otolaryngol*, 2022.

[91] E. Harvey *et al.*, “Impact of Demographics and Clinical Features on Initial Treatment Pathway for Vestibular Schwannoma,” *Otology & Neurotology*, vol. 43, no. 9, pp. 1078–1084, 2022.

[92] A. Thomas *et al.*, “The impact of social determinants of health and clinical comorbidities on post-tympanotomy tube otorrhea,” *Int J Pediatr Otorhinolaryngol*, p. 110986, Nov. 2021, doi: 10.1016/J.IJPORL.2021.110986.

[93] M. E. Tinetti, M. Speechley, and S. F. Ginter, “Risk factors for falls among elderly persons living in the community,” *New England journal of medicine*, vol. 319, no. 26, pp. 1701–1707, 1988.

[94] D. Schoene, C. Heller, Y. N. Aung, C. C. Sieber, W. Kemmler, and E. Freiberger, “A systematic review on the influence of fear of falling on quality of life in older people: is there a role for falls?,” *Clin Interv Aging*, pp. 701–719, 2019.

[95] L. Z. Rubenstein, K. R. Josephson, and A. S. Robbins, “Falls in the nursing home,” *Ann Intern Med*, vol. 121, no. 6, pp. 442–451, 1994.

[96] M. E. Tinetti, M. Speechley, and S. F. Ginter, “Risk factors for falls among elderly persons living in the community,” *New England journal of medicine*, vol. 319, no. 26, pp. 1701–1707, 1988.

[97] M. Pirrie, G. Saini, R. Angeles, F. Marzanek, J. Parascandalo, and G. Agarwal, “Risk of falls and fear of falling in older adults residing in public housing in Ontario, Canada: findings from a multisite observational study,” *BMC Geriatr*, vol. 20, no. 1, pp. 1–8, 2020.

[98] S. Zhao *et al.*, “Population ageing and injurious falls among one million elderly people who used emergency medical services from 2010 to 2017 in Beijing, China: a longitudinal observational study,” *BMJ Open*, vol. 9, no. 6, p. e028292, 2019.

[99] T. Kim, S. D. Choi, and S. Xiong, “Epidemiology of fall and its socioeconomic risk factors in community-dwelling Korean elderly,” *PLoS One*, vol. 15, no. 6, p. e0234787, 2020.

[100] C. W. T. Lo *et al.*, “Acceptability and feasibility of a community-based strength, balance, and Tai Chi rehabilitation program in improving physical function and balance of patients after total knee arthroplasty: study protocol for a pilot randomized controlled trial,” *Trials*, vol. 22, no. 1, pp. 1–11, 2021.

[101] K. Fiscella and P. Shin, “The inverse care law: implications for healthcare of vulnerable populations,” *J Ambul Care Manage*, vol. 28, no. 4, pp. 304–312, 2005.

[102] R. Agarwal and M. G. Gopinath, *A proposal to end the COVID-19 pandemic*. International Monetary Fund, 2021.

[103] M. Kamruzzaman and M. A. Hakim, “Socio-economic status of slum dwellers: an empirical study on the capital city of Bangladesh,” *Age (years)*, vol. 20, no. 35, p. 11, 2016.

[104] M. L. C. Valencia, B. T. Tran, M. K. Lim, K. S. Choi, and J.-K. Oh, “Association between socioeconomic status and early initiation of smoking, alcohol drinking, and sexual behavior among Korean adolescents,” *Asia Pacific Journal of Public Health*, vol. 31, no. 5, pp. 443–453, 2019.

[105] W. J. Smith *et al.*, “Promoting physical activity in rural settings: effectiveness and potential strategies,” *Transl J Am Coll Sports Med*, vol. 6, no. 4, p. e000180, 2021.

[106] B. Kaplan and S. Litewka, “Ethical challenges of telemedicine and telehealth,” *Cambridge Quarterly of Healthcare Ethics*, vol. 17, no. 4, pp. 401–416, 2008.

[107] S. Hopewell *et al.*, “Multifactorial and multiple component interventions for preventing falls in older people living in the community,” *Cochrane database of systematic reviews*, no. 7, 2018.

[108] J. Choi *et al.*, “Development and validation of the fall risk perception questionnaire for patients in acute care hospitals,” *J Clin Nurs*, vol. 30, no. 3–4, pp. 406–414, 2021.

[109] M.-H. Tseng and H.-C. Wu, “Integrating socioeconomic status and spatial factors to improve the accessibility of community care resources using maximum-equity optimization of supply capacity allocation,” *Int J Environ Res Public Health*, vol. 18, no. 10, p. 5437, 2021.

[110] S. Keesara, A. Jonas, and K. Schulman, “Covid-19 and health care’s digital revolution,” *New England Journal of Medicine*, vol. 382, no. 23, p. e82, 2020.

[111] A. G. Ekeland, A. Bowes, and S. Flottorp, “Effectiveness of telemedicine: a systematic review of reviews,” *Int J Med Inform*, vol. 79, no. 11, pp. 736–771, 2010.

[112] X. Zhou *et al.*, “The Role of Telehealth in Reducing the Mental Health Burden from COVID-19,” *Telemedicine and e-Health*, vol. 26, no. 4. Mary Ann Liebert Inc., pp. 377–379, Apr. 01, 2020. doi: 10.1089/tmj.2020.0068.

[113] M. E. Reed *et al.*, “Patient characteristics associated with choosing a telemedicine visit vs office visit with the same primary care clinicians,” *JAMA Netw Open*, vol. 3, no. 6, pp. e205873–e205873, 2020.

[114] C. M. Castro Sweet *et al.*, “Outcomes of a digital health program with human coaching for diabetes risk reduction in a Medicare population,” *J Aging Health*, vol. 30, no. 5, pp. 692–710, 2018.

[115] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, “Big data in health care: using analytics to identify and manage high-risk and high-cost patients,” *Health Aff*, vol. 33, no. 7, pp. 1123–1131, 2014.

[116] P.-H. Chen, H. Zafar, M. Galperin-Aizenberg, and T. Cook, “Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports,” *J Digit Imaging*, vol. 31, pp. 178–184, 2018.

[117] A. W. Forsyth *et al.*, “Machine learning methods to extract documentation of breast cancer symptoms from electronic health records,” *J Pain Symptom Manage*, vol. 55, no. 6, pp. 1492–1499, 2018.

[118] V. Okunrintemi *et al.*, “Association of Income Disparities with Patient-Reported Healthcare Experience,” *J Gen Intern Med*, vol. 34, no. 6, pp. 884–892, Jun. 2019, doi: 10.1007/s11606-019-04848-4.

[119] N. Houssami, G. Kirkpatrick-Jones, N. Noguchi, and C. I. Lee, “Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI’s potential in breast screening practice,” *https://doi.org/10.1080/17434440.2019.1610387*, vol. 16, no. 5, pp. 351–362, May 2019, doi: 10.1080/17434440.2019.1610387.

[120] F. De Carlo *et al.*, “Scientific data exchange: a schema for HDF5-based storage of raw and analyzed data,” *J Synchrotron Radiat*, vol. 21, no. 6, pp. 1224–1230, 2014.

[121] G. Mujtaba *et al.*, “Clinical text classification research trends: Systematic literature review and open issues,” *Expert Syst Appl*, vol. 116, pp. 494–520, Feb. 2019, doi: 10.1016/J.ESWA.2018.09.034.

[122]D. B. Aronow, F. Fangfang, and W. B. Croft, “Ad Hoc Classification of Radiology Reports,” *Journal of the American Medical Informatics Association*, vol. 6, no. 5, pp. 393–411, Sep. 1999, doi: 10.1136/JAMIA.1999.0060393.

[123] B. J. Thomas, H. Ouellette, E. F. Halpern, and D. I. Rosenthal, “Automated Computer-Assisted Categorization of Radiology Reports,” *American Journal or Roentgenology,* vol. 184, no. 2, pp. 687–690, Nov. 2005, doi: 10.2214/AJR.184.2.01840687.

[124] J. Luo, C. Erbe, and D. R. Friedland, “Unique clinical language patterns among expert vestibular providers can predict vestibular diagnoses,” *Otology and Neurotology*, vol. 39, no. 9, pp. 1163–1171, 2018, doi: 10.1097/MAO.0000000000001930.

[125] D. D. Lewis, “A Comparison of Two Learning Algorithms for Text Categorization 1 Introduction 2 Text Categorization : Nature and Approaches,” *In Proceeding of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR`94)*, vol. 33, pp. 1–14, 1994.

[126] B. Raja Srinivasa Reddy and B. B. Kadaru, “An integrated hybrid feature selection based ensemble learning model for parkinson and alzheimer’s disease prediction,” *International Journal of Applied Engineering Research*, vol. 12, no. 22, pp. 11989–12003, 2017, Accessed: Apr. 16, 2020. [Online]. Available: http://www.ripublication.com

[127] B. de Bruijn, A. Cranney, S. O’Donnell, J. D. Martin, and A. J. Forster, “Identifying Wrist Fracture Patients with High Accuracy by Automatic Categorization of X-ray Reports,” *Journal of the American Medical Informatics Association*, vol. 13, no. 6, pp. 696–698, Nov. 2006, doi: 10.1197/JAMIA.M1995.

[128] A. McCallum and K. Nigam, “A comparison of event models for naive bayes text classification,” *Association for the Advancement of Artificial Intelligence*, vol. 752, no. 1, pp. 41–48, 1998, Accessed: Jul. 14, 2021. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.9324&rep=rep1&type=pdf

[129] K. M. Schneider, “Techniques for improving the performance of naive bayes for text classification,” in *Lecture Notes in Computer Science*, 2005, pp. 682–693. doi: 10.1007/978-3-540-30586-6_76.

[130] Y. Wang *et al.*, “A clinical text classification paradigm using weak supervision and deep representation,” *BMC Medical Informatics and Decision Making 2019 19:1*, vol. 19, no. 1, pp. 1–13, Jan. 2019, doi: 10.1186/S12911-018-0723-6.

[131] Y. P. Qin and X. K. Wang, “Study on multi-label text classification based on SVM,” *6th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2009*, vol. 1, pp. 300–304, 2009, doi: 10.1109/FSKD.2009.207.

[132] G. Zuccon *et al.*, “Automatic Classification of Free-Text Radiology Reports to Identify Limb Fractures using Machine Learning and the SNOMED CT Ontology,” *AMIA Summits on Translational Science Proceedings*, vol. 2013, p. 300, 2013, Accessed: Jul. 14, 2021. [Online]. Available: /pmc/articles/PMC3845773/

[133] T. Joachims, “Text categorization with Support Vector Machines: Learning with many relevant features,” *European Conference on Machine Learning*, pp. 137–142, 1998, doi: 10.1007/BFB0026683.

[134] V. Chaurasia and S. Pal, “Data Mining Approach to Detect Heart Diseases,” *International Journal of Advanced Computer Science and Information Technology*, vol. 2, no. 4, pp. 56–66, 2014, Accessed: Apr. 16, 2020. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2376653

[135] P. Vateekul and M. Kubat, “Fast induction of multiple decision trees in text categorization from large scale, imbalanced, and multi-label data,” *ICDM Workshops 2009 - IEEE International Conference on Data Mining*, pp. 320–325, 2009, doi: 10.1109/ICDMW.2009.94.

[136] D. E. Johnson, F. J. Oles, T. Zhang, and T. Goetz, “A decision-tree-based symbolic rule induction system for text categorization,” *IBM Systems Journal*, vol. 41, no. 3, pp. 428–437, 2002, doi: 10.1147/SJ.413.0428.

[137] H. Freitas Da Cruz, F. Schneider, and M.-P. Schapranow, “Prediction of Acute Kidney Injury in Cardiac Surgery Patients: Interpretation using Local Interpretable Model-agnostic Explanations,” *HEALTHINF*, pp. 380–387, 2019, doi: 10.5220/0007399203800387.

[138] Z. Dai, Z. Li, and L. Han, “Bonebert: A bert-based automated information extraction system of radiology reports for bone fracture detection and diagnosis,” in *Advances in Intelligent Data Analysis XIX: 19th International Symposium on Intelligent Data Analysis, IDA 2021, Porto, Portugal, April 26–28, 2021, Proceedings 19*, 2021, pp. 263–274.

[139] E. S. Kayi, K. Yadav, J. M. Chamberlain, and H.-A. Choi, “Topic modeling for classification of clinical reports,” *arXiv preprint arXiv:1706.06177*, 2017.

[140] Dipanjan Sarkar, “The Importance of Human Interpretable Machine Learning,” *Towards Data Science*, May 24, 2018. https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-need-and-importance-of-model-interpretation-2ed758f5f476 (accessed Jul. 14, 2021).

[141] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable AI: A Review of Machine Learning Interpretability Methods,” *Entropy 2021, Vol. 23, Page 18*, vol. 23, no. 1, p. 18, Dec. 2020, doi: 10.3390/E23010018.

[142] C. Molnar, G. Casalicchio, and B. Bischl, “Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges,” *Communications in Computer and Information Science*, vol. 1323, pp. 417–431, Sep. 2020, doi: 10.1007/978-3-030-65965-3_28.

[143] C. Molnar, *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*. Leanupb, 2019. Accessed: Jul. 14, 2021. [Online]. Available: http://leanpub.com/interpretable-machine-learning

[144] Scott Lundberg, “SHAP documentation,” 2018. https://shap.readthedocs.io/en/latest/index.html (accessed Jul. 14, 2021).

[145] Mikahail Korobov and Konstantin Lopuhin, “ELI5 documentation,” 2017. https://eli5.readthedocs.io/en/latest/index.html (accessed Jul. 14, 2021).

[146] InterpretML Team, “InterpretML documentation,” 2021. https://interpret.ml/docs/intro.html (accessed Jul. 14, 2021).

[147] J. Allyn *et al.*, “A comparison of a machine learning model with EuroSCORE II in predicting mortality after elective cardiac surgery: A decision curve analysis,” *PLoS One*, vol. 12, no. 1, 2017, doi: 10.1371/journal.pone.0169772.

[148] A. E. U. Cerna *et al.*, “Interpretable Neural Networks for Predicting Mortality Risk using Multi-modal Electronic Health Records,” *ArXiv*, Jan. 2019, Accessed: Jul. 14, 2021. [Online]. Available: https://eugdpr.org/

[149] I. Neves *et al.*, “Interpretable heartbeat classification using local model-agnostic explanations on ECGs,” *Comput Biol Med*, vol. 133, p. 104393, Jun. 2021, doi: 10.1016/J.COMPBIOMED.2021.104393.

[150] M. Jovanovic, S. Radovanovic, M. Vukicevic, S. Van Poucke, and B. Delibasic, “Building interpretable predictive models for pediatric hospital readmission using Tree-Lasso logistic regression,” *Artif Intell Med*, vol. 72, pp. 12–21, 2016.

[151] R. Gargeya and T. Leng, “Automated Identification of Diabetic Retinopathy Using Deep Learning,” *Ophthalmology*, vol. 124, no. 7, pp. 962–969, Jul. 2017, doi: 10.1016/j.ophtha.2017.02.008.

[152] A. J. H. Kind and W. R. Buckingham, “Making neighborhood-disadvantage metrics accessible—the neighborhood atlas,” *N Engl J Med*, vol. 378, no. 26, p. 2456, 2018.

[153] M. A. Butler, *Rural-urban continuum codes for metro and nonmetro counties*. US Department of Agriculture, Economic Research Service, Agriculture, 1990.

[154] D. F. López-Cevallos and C. Chi, “Assessing the context of health care utilization in Ecuador: A spatial and multilevel analysis,” *BMC Health Serv Res*, vol. 10, no. 1, pp. 1–10, 2010.

[155] G. Trends, “Public health and aging: trends in aging—United States and worldwide,” *Public Health*, vol. 347, pp. 921–925, 2003.

[156] S. Colby and J. M. Ortman, *The baby boom cohort in the United States: 2012 to 2060*. US Department of Commerce, Economics and Statistics Administration, US~…, 2014.

[157] S. G. Bromfield *et al.*, “Blood pressure, antihypertensive polypharmacy, frailty, and risk for serious fall injuries among older treated adults with hypertension,” *Hypertension*, vol. 70, no. 2, pp. 259–266, 2017.

[158] D. A. Lawlor, R. Patel, and S. Ebrahim, “Association between falls in elderly women and chronic diseases and drug use: cross sectional study,” *Bmj*, vol. 327, no. 7417, pp. 712–717, 2003.

[159] L. A. Cooper, M. N. Hill, and N. R. Powe, “Designing and evaluating interventions to eliminate racial and ethnic disparities in health care,” *J Gen Intern Med*, vol. 17, no. 6, pp. 477–486, 2002.

[160] J. Johansson, A. Nordström, and P. Nordström, “Greater fall risk in elderly women than in men is associated with increased gait variability during multitasking,” *J Am Med Dir Assoc*, vol. 17, no. 6, pp. 535–540, 2016.

[161] M. L. Callisaya *et al.*, “Gait, gait variability and the risk of multiple incident falls in older people: a population-based study,” *Age Ageing*, vol. 40, no. 4, pp. 481–487, 2011.

[162] W. H. Frey and D. Myers, “Racial segregation in US metropolitan areas and cities, 1990–2000: Patterns, trends, and explanations,” *Population studies center research report*, vol. 5, p. 573, 2005.

[163] R. Vijayakrishnan *et al.*, “Prevalence of Heart Failure Signs and Symptoms in a Large Primary Care Population Identified Through the Use of Text and Data Mining of the Electronic Health Record,” *J Card Fail*, vol. 20, no. 7, pp. 459–464, Jul. 2014, doi: 10.1016/J.CARDFAIL.2014.03.008.

[164] D. R. Friedland, S. Tarima, C. Erbe, and A. Miles, “Development of a Statistical Model for the Prediction of Common Vestibular Diagnoses,” *JAMA Otolaryngology–Head & Neck Surgery*, vol. 142, no. 4, pp. 351–356, Apr. 2016, doi: 10.1001/JAMAOTO.2015.3663.

[165] A. Mehrotra *et al.*, “Applying a natural language processing tool to electronic health records to assess performance on colonoscopy quality measures,” *Gastrointest Endosc*, vol. 75, no. 6, pp. 1233-1239.e14, Jun. 2012, doi: 10.1016/J.GIE.2012.01.045.

[166] S. Ghafouri-Fard, M. Taheri, M. D. Omrani, A. Daaee, H. Mohammad-Rahimi, and H. Kazazi, “Application of Single-Nucleotide Polymorphisms in the Diagnosis of Autism Spectrum Disorders: A Preliminary Study with Artificial Neural Networks,” *Journal of*

*Molecular Neuroscience 2019 68:4*, vol. 68, no. 4, pp. 515–521, Apr. 2019, doi: 10.1007/S12031-019-01311-1.

[167] I. P. de Sousa, M. M. B. R. Vellasco, and E. C. da Silva, "Local Interpretable Model-Agnostic Explanations for Classification of Lymph Node Metastases," *Sensors (Basel)*, vol. 19, no. 13, Jul. 2019, doi: 10.3390/S19132969.

[168] S. Kulshrestha *et al.*, "Comparison and interpretability of machine learning models to predict severity of chest injury," *JAMIA Open*, vol. 4, no. 1, pp. 1–8, Mar. 2021, doi: 10.1093/JAMIAOPEN/OOAB015.

[169] K Donnelly, "SNOMED-CT: The advanced terminology and coding system for eHealth," *Stud Health Technol Inform*, vol. 121, pp. 279–279, 2006.

# 7. Appendices

## Appendix A. List of 126 Zip Codes and Utilization Rates For Fall And Non-Fall Cohorts

| | fall | | | fall | | | fall | |
|---|---|---|---|---|---|---|---|---|
| zip | UR | Non-fall UR | zip | UR | Non-fall UR | zip | UR | Non-fall UR |
| 53002 | 3.8% | 26.8% | 53119 | 0.9% | 12.6% | 53192 | 4.0% | 46.0% |
| 53004 | 0.5% | 9.6% | 53120 | 1.2% | 14.6% | 53195 | 0.0% | 9.3% |
| 53005 | 5.8% | 37.6% | 53121 | 0.4% | 8.0% | 53202 | 2.5% | 18.8% |
| 53007 | 8.3% | 44.6% | 53122 | 11.0% | 48.6% | 53203 | 2.3% | 21.3% |
| 53012 | 1.5% | 18.7% | 53125 | 1.1% | 13.8% | 53204 | 1.1% | 10.1% |
| 53017 | 4.7% | 36.7% | 53126 | 1.1% | 14.8% | 53205 | 3.4% | 21.9% |
| 53018 | 1.2% | 20.5% | 53128 | 0.2% | 3.8% | 53206 | 3.1% | 23.7% |
| 53021 | 1.6% | 14.1% | 53129 | 3.9% | 25.9% | 53207 | 1.2% | 12.7% |
| 53022 | 6.5% | 36.5% | 53130 | 4.4% | 26.6% | 53208 | 3.3% | 20.8% |
| 53024 | 1.1% | 16.1% | 53132 | 2.1% | 18.2% | 53209 | 3.3% | 22.8% |
| 53027 | 3.5% | 25.2% | 53137 | 0.4% | 8.0% | 53210 | 2.8% | 18.0% |
| 53029 | 1.6% | 19.5% | 53139 | 1.2% | 15.0% | 53211 | 1.8% | 15.2% |
| 53033 | 5.6% | 41.3% | 53140 | 0.9% | 12.3% | 53212 | 1.8% | 15.2% |
| 53036 | 0.6% | 9.8% | 53142 | 0.7% | 11.6% | 53213 | 8.6% | 30.3% |
| 53037 | 6.0% | 31.5% | 53143 | 0.7% | 10.5% | 53214 | 4.8% | 23.7% |
| 53038 | 0.2% | 4.4% | 53144 | 0.7% | 9.8% | 53215 | 1.0% | 9.1% |
| 53040 | 6.4% | 35.3% | 53146 | 2.9% | 22.8% | 53216 | 3.0% | 20.3% |
| 53045 | 5.3% | 36.0% | 53147 | 0.5% | 9.6% | 53217 | 2.4% | 24.6% |
| 53046 | 7.1% | 42.8% | 53149 | 1.0% | 14.0% | 53218 | 2.9% | 18.1% |
| 53051 | 8.7% | 45.2% | 53150 | 2.7% | 20.5% | 53219 | 2.9% | 18.7% |
| 53058 | 1.5% | 19.9% | 53151 | 4.5% | 29.6% | 53220 | 2.8% | 20.0% |
| 53066 | 1.3% | 17.0% | 53153 | 0.7% | 17.3% | 53221 | 1.9% | 15.3% |
| 53069 | 2.5% | 26.8% | 53154 | 1.4% | 13.0% | 53222 | 6.1% | 26.2% |
| 53072 | 2.5% | 25.9% | 53156 | 0.6% | 9.8% | 53223 | 4.6% | 26.9% |
| 53074 | 0.9% | 12.9% | 53158 | 0.7% | 10.8% | 53224 | 4.4% | 23.4% |
| 53076 | 5.0% | 35.0% | 53167 | 0.8% | 8.8% | 53225 | 6.3% | 26.7% |
| 53080 | 1.1% | 13.8% | 53168 | 0.6% | 9.6% | 53226 | 13.3% | 45.1% |
| 53086 | 4.5% | 28.8% | 53170 | 0.5% | 8.9% | 53227 | 4.5% | 24.8% |
| 53089 | 4.1% | 29.8% | 53172 | 1.3% | 13.0% | 53228 | 5.1% | 29.2% |
| 53090 | 5.5% | 28.5% | 53177 | 0.8% | 11.2% | 53233 | 1.4% | 9.4% |
| 53092 | 2.5% | 29.7% | 53178 | 0.6% | 12.2% | 53235 | 1.8% | 15.6% |
| 53094 | 0.4% | 6.0% | 53179 | 0.4% | 5.5% | 53295 | 3.6% | 35.1% |
| 53095 | 9.7% | 44.4% | 53181 | 0.6% | 7.7% | 53402 | 1.1% | 15.4% |
| 53097 | 1.6% | 19.9% | 53182 | 1.4% | 15.0% | 53403 | 0.7% | 12.1% |
| 53103 | 1.5% | 18.6% | 53183 | 1.5% | 19.3% | 53404 | 0.6% | 9.4% |
| 53104 | 0.8% | 11.7% | 53184 | 0.3% | 4.8% | 53405 | 0.8% | 13.4% |
| 53105 | 0.8% | 11.6% | 53185 | 1.4% | 15.0% | 53406 | 1.2% | 17.3% |
| 53108 | 1.3% | 18.2% | 53186 | 1.6% | 20.1% | 53538 | 0.1% | 1.8% |
| 53110 | 1.4% | 13.9% | 53188 | 1.3% | 17.4% | 53549 | 0.3% | 3.0% |
| 53114 | 0.1% | 3.6% | 53189 | 1.3% | 15.1% | 53551 | 0.1% | 2.6% |
| 53115 | 0.3% | 6.1% | 53190 | 0.1% | 2.0% | 53585 | 0.3% | 2.8% |
| 53118 | 1.3% | 19.3% | 53191 | 0.5% | 10.5% | 53594 | 0.1% | 1.1% |

UR = Utilization rate.

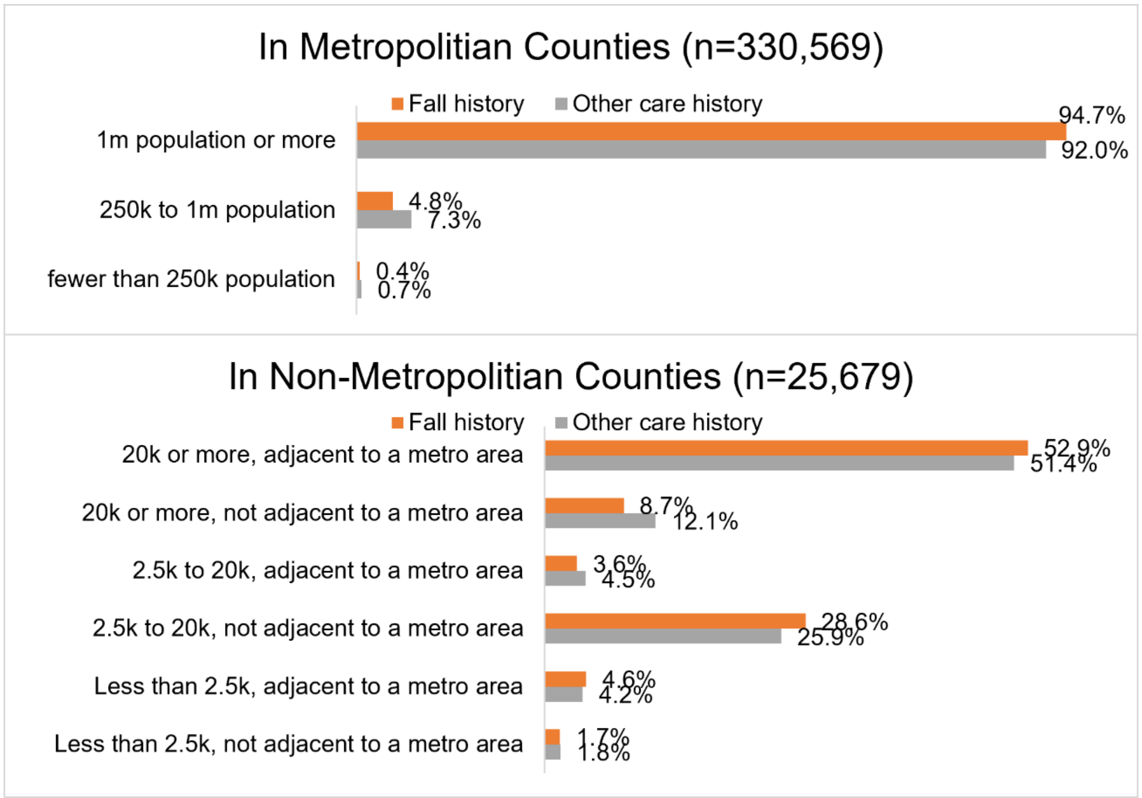
## Appendix B. Socioeconomic Variables for the 126 Zip Codes

| Zip | MHI | CER | white | PIR | zip | MHI | CER | white | PIR | Zip | MHI | CER | white | PIR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 53002 | 85478 | 25% | 98% | 60% | 53119 | 87553 | 29% | 96% | 21% | 53192 | 63293 | 0% | 100% | 74% |
| 53004 | 77989 | 31% | 94% | 19% | 53120 | 66815 | 30% | 95% | 19% | 53195 | 52125 | 13% | 78% | 6% |
| 53005 | 97202 | 57% | 90% | 54% | 53121 | 65065 | 30% | 95% | 10% | 53202 | 52441 | 69% | 83% | 34% |
| 53007 | 41925 | 25% | 91% | 47% | 53122 | 113750 | 71% | 95% | 62% | 53203 | 66161 | 63% | 81% | 57% |
| 53012 | 88844 | 52% | 95% | 29% | 53125 | 78971 | 58% | 97% | 10% | 53204 | 28218 | 8% | 37% | 10% |
| 53017 | 106577 | 40% | 96% | 65% | 53126 | 77393 | 30% | 96% | 22% | 53205 | 23125 | 10% | 5% | 13% |
| 53018 | 97482 | 53% | 94% | 27% | 53128 | 56057 | 19% | 95% | 6% | 53206 | 22877 | 7% | 2% | 11% |
| 53021 | 82750 | 19% | 94% | 26% | 53129 | 64622 | 40% | 89% | 32% | 53207 | 58646 | 37% | 87% | 23% |
| 53022 | 79230 | 39% | 91% | 67% | 53130 | 71393 | 42% | 94% | 41% | 53208 | 31592 | 26% | 35% | 26% |
| 53024 | 75592 | 45% | 94% | 23% | 53132 | 76548 | 39% | 84% | 28% | 53209 | 33395 | 21% | 24% | 20% |
| 53027 | 68701 | 26% | 96% | 48% | 53137 | 75893 | 24% | 98% | 11% | 53210 | 34516 | 20% | 21% | 23% |
| 53029 | 91435 | 47% | 96% | 37% | 53139 | 78831 | 30% | 96% | 19% | 53211 | 60195 | 70% | 85% | 25% |
| 53033 | 92620 | 35% | 98% | 67% | 53140 | 38486 | 18% | 81% | 8% | 53212 | 33597 | 31% | 41% | 16% |
| 53036 | 75179 | 31% | 96% | 17% | 53142 | 66607 | 34% | 85% | 12% | 53213 | 73308 | 60% | 87% | 68% |
| 53037 | 74327 | 30% | 98% | 69% | 53143 | 45286 | 19% | 78% | 8% | 53214 | 44343 | 23% | 82% | 34% |
| 53038 | 75000 | 30% | 96% | 8% | 53144 | 55414 | 25% | 81% | 10% | 53215 | 35803 | 10% | 65% | 11% |
| 53040 | 70938 | 21% | 96% | 66% | 53146 | 83586 | 33% | 96% | 41% | 53216 | 34977 | 18% | 11% | 21% |
| 53045 | 100438 | 63% | 86% | 52% | 53147 | 60943 | 35% | 94% | 8% | 53217 | 100262 | 71% | 87% | 38% |
| 53046 | 56771 | 19% | 98% | 70% | 53149 | 86899 | 34% | 95% | 23% | 53218 | 37692 | 13% | 13% | 20% |
| 53051 | 76944 | 42% | 89% | 67% | 53150 | 85744 | 36% | 97% | 40% | 53219 | 52468 | 22% | 85% | 27% |
| 53058 | 92792 | 51% | 96% | 27% | 53151 | 76772 | 43% | 92% | 49% | 53220 | 48523 | 24% | 85% | 27% |
| 53066 | 86343 | 44% | 96% | 22% | 53153 | 90250 | 33% | 96% | 25% | 53221 | 50645 | 22% | 81% | 18% |
| 53069 | 115568 | 54% | 94% | 30% | 53154 | 70530 | 33% | 86% | 27% | 53222 | 52811 | 39% | 64% | 46% |
| 53072 | 82032 | 49% | 94% | 39% | 53156 | 57989 | 20% | 94% | 12% | 53223 | 45973 | 28% | 36% | 28% |
| 53074 | 64446 | 32% | 97% | 17% | 53158 | 76851 | 35% | 90% | 13% | 53224 | 46766 | 25% | 32% | 35% |
| 53076 | 103000 | 36% | 98% | 73% | 53167 | 51579 | 19% | 100% | 9% | 53225 | 42665 | 19% | 34% | 34% |
| 53080 | 59545 | 23% | 95% | 19% | 53168 | 74088 | 24% | 96% | 12% | 53226 | 71121 | 57% | 87% | 69% |
| 53086 | 78796 | 33% | 98% | 64% | 53170 | 61228 | 25% | 96% | 10% | 53227 | 48613 | 26% | 86% | 34% |
| 53089 | 87106 | 38% | 95% | 61% | 53172 | 51484 | 21% | 91% | 19% | 53228 | 61839 | 33% | 88% | 40% |
| 53090 | 66385 | 23% | 96% | 58% | 53177 | 64920 | 17% | 83% | 12% | 53233 | 14920 | 23% | 55% | 11% |
| 53092 | 96627 | 64% | 94% | 37% | 53178 | 60298 | 19% | 97% | 12% | 53235 | 41719 | 32% | 89% | 20% |
| 53094 | 49514 | 21% | 95% | 7% | 53179 | 68199 | 22% | 97% | 7% | 53295 | 63293 | 10% | 72% | 8% |
| 53095 | 63157 | 30% | 95% | 63% | 53181 | 65168 | 21% | 98% | 9% | 53402 | 58134 | 28% | 83% | 17% |
| 53097 | 107667 | 62% | 89% | 39% | 53182 | 66572 | 19% | 93% | 16% | 53403 | 43411 | 21% | 62% | 10% |
| 53103 | 86809 | 26% | 97% | 29% | 53183 | 99095 | 49% | 96% | 28% | 53404 | 35013 | 13% | 50% | 8% |
| 53104 | 71462 | 23% | 97% | 13% | 53184 | 51377 | 27% | 93% | 5% | 53405 | 52978 | 20% | 74% | 13% |
| 53105 | 66699 | 25% | 96% | 14% | 53185 | 86729 | 31% | 96% | 25% | 53406 | 64478 | 31% | 82% | 15% |
| 53108 | 68673 | 17% | 95% | 23% | 53186 | 57423 | 36% | 88% | 28% | 53538 | 55340 | 22% | 93% | 2% |
| 53110 | 49882 | 21% | 91% | 18% | 53188 | 66359 | 36% | 90% | 26% | 53549 | 54603 | 18% | 95% | 4% |
| 53114 | 63190 | 17% | 96% | 4% | 53189 | 87062 | 40% | 92% | 28% | 53551 | 69227 | 36% | 93% | 4% |
| 53115 | 53868 | 19% | 92% | 5% | 53190 | 43339 | 34% | 90% | 2% | 53585 | 50365 | 12% | 90% | 3% |
| 53118 | 79577 | 34% | 96% | 21% | 53191 | 63190 | 41% | 95% | 8% | 53594 | 75306 | 22% | 94% | 1% |

MHI = Median Household Income; CER = College educated Rate; PIR = Privately insured Rate.
Data source is from U.S. Census Bureau Releases of the 2014-2018 American Community Survey (ACS) 5-year estimates of the social, housing and demographic information.

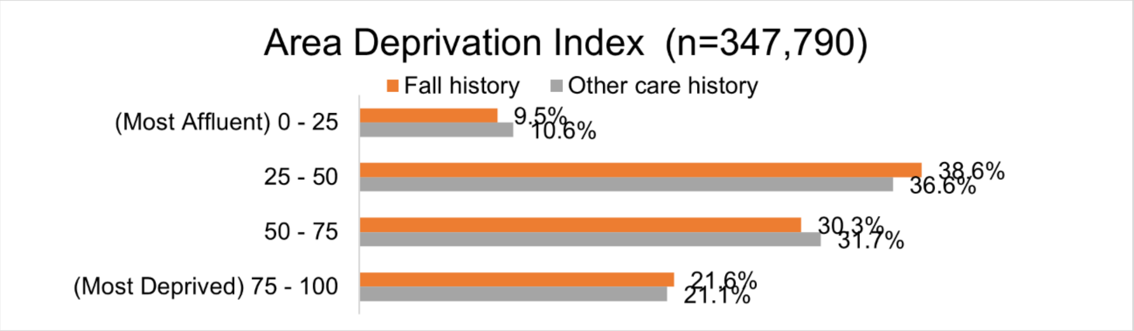
## Appendix C. Summary of Rural-Urban Continuum Codes and Cohort Groups

| **Metropolitan Counties*** | | Patient with Fall history | | Patient with Fall history | |
|---|---|---|---|---|---|
| | Number of patients | 66357 | | 428684 | |
| **Metropolitan Counties** | | | | | |
| Code | Description | | | | |
| 1 | 1m population or more | 45483 | 94.7% | 264089 | 92.0% |
| 2 | 250k to 1m population | 2321 | 4.8% | 21061 | 7.3% |
| 3 | fewer than 250k population | 210 | 0.4% | 1905 | 0.7% |
| **Nonmetropolitan Counties** | | | | | |
| Code | Description | | | | |
| 4 | 20k or more, adjacent to a metro area | 878 | 52.9% | 12356 | 51.4% |
| 5 | 20k or more, not adjacent to a metro area | 144 | 8.7% | 2914 | 12.1% |
| 6 | 2.5k to 20k, adjacent to a metro area | 59 | 3.6% | 1086 | 4.5% |
| 7 | 2.5k to 20k, not adjacent to a metro area | 474 | 28.6% | 6231 | 25.9% |
| 8 | Less than 2.5k, adjacent to a metro area | 76 | 4.6% | 1010 | 4.2% |
| 9 | Less than 2.5k, not adjacent to a metro area | 28 | 1.7% | 423 | 1.8% |



## Appendix D. Quartile Categorization of Area Deprivation Index

| **Area Deprivation Index** | Patient with Fall history (%) | | Patient without Fall history (%) | |
|---|---|---|---|---|
| (Most Affluent) 0 - 25 | 4487 | 9.5% | 31705 | 10.6% |
| 25 - 50 | 18256 | 38.6% | 110056 | 36.6% |
| 50 - 75 | 14340 | 30.3% | 95220 | 31.7% |
| (Most Deprived) 75 - 100 | 10215 | 21.6% | 63511 | 21.1% |



## Appendix E: Top 40 Word Frequency Gaps in Positive And Negative Word Counts

| | Word | frequency in positive reports | frequency in negative reports | weighted difference | word | frequency in positive reports | frequency in negative reports | weighted difference |
|---|---|---|---|---|---|---|---|---|
| 1 | fracture | 8.8 | 0.5 | 8.28 | calcification | 0.1 | 0.8 | -0.70 |
| 2 | left | 9.6 | 2.8 | 6.79 | none | 0.5 | 1.1 | -0.63 |
| 3 | bone | 7.5 | 2.0 | 5.49 | mild | 0.3 | 0.9 | -0.61 |
| 4 | temporal | 6.9 | 1.5 | 5.48 | lung | 0.1 | 0.7 | -0.60 |
| 5 | right | 8.5 | 3.3 | 5.16 | lesion | 0.1 | 0.6 | -0.53 |
| 6 | canal | 6.6 | 3.0 | 3.56 | loss | 0.0 | 0.6 | -0.53 |
| 7 | mastoid | 3.6 | 1.4 | 2.27 | thickening | 0.2 | 0.7 | -0.50 |
| 8 | head | 2.6 | 0.5 | 2.10 | normal | 3.6 | 4.1 | -0.49 |
| 9 | ear | 3.8 | 1.7 | 2.07 | hearing | 0.1 | 0.5 | -0.48 |
| 10 | air | 3.2 | 1.2 | 2.05 | cm | 0.1 | 0.4 | -0.28 |
| 11 | ct | 4.9 | 2.9 | 1.95 | chronic | 0.1 | 0.4 | -0.27 |
| 12 | within | 2.8 | 0.8 | 1.93 | disc | 0.1 | 0.3 | -0.27 |
| 13 | facial | 2.1 | 0.2 | 1.90 | dose | 0.4 | 0.7 | -0.27 |
| 14 | hemorrhage | 1.9 | 0.1 | 1.79 | clinical | 0.8 | 1.0 | -0.26 |
| 15 | middle | 3.0 | 1.4 | 1.65 | narrowing | 0.0 | 0.3 | -0.26 |
| 16 | auditory | 3.2 | 1.6 | 1.65 | disease | 0.2 | 0.4 | -0.26 |
| 17 | fossa | 2.1 | 0.6 | 1.54 | change | 0.6 | 0.9 | -0.25 |
| 18 | anterior | 1.6 | 0.2 | 1.38 | dehiscence | 0.1 | 0.4 | -0.24 |
| 19 | sinus | 2.4 | 1.0 | 1.37 | unremarkable | 1.3 | 1.5 | -0.24 |
| 20 | sphenoid | 1.5 | 0.1 | 1.36 | mucosal | 0.2 | 0.4 | -0.24 |
| 21 | external | 2.2 | 0.8 | 1.35 | contrast | 1.2 | 1.4 | -0.23 |
| 22 | carotid | 2.4 | 1.0 | 1.33 | well | 0.9 | 1.1 | -0.21 |
| 23 | cell | 2.4 | 1.1 | 1.30 | focal | 0.1 | 0.3 | -0.21 |
| 24 | cavity | 1.8 | 0.5 | 1.29 | effusion | 0.2 | 0.4 | -0.19 |
| 25 | posterior | 1.7 | 0.4 | 1.27 | parenchyma | 0.0 | 0.2 | -0.18 |
| 26 | nondisplaced | 1.2 | 0.0 | 1.22 | history | 0.3 | 0.5 | -0.18 |
| 27 | seen | 1.8 | 0.7 | 1.14 | atherosclerotic | 0.1 | 0.3 | -0.18 |
| 28 | capsule | 1.2 | 0.1 | 1.10 | technique | 1.9 | 2.1 | -0.18 |
| 29 | otic | 1.2 | 0.1 | 1.09 | reduction | 0.4 | 0.6 | -0.18 |
| 30 | nerve | 1.3 | 0.2 | 1.07 | high | 0.1 | 0.2 | -0.18 |
| 31 | extends | 1.1 | 0.1 | 1.06 | patient | 0.9 | 1.0 | -0.18 |
| 32 | extending | 1.1 | 0.1 | 1.03 | reconstruction | 0.4 | 0.6 | -0.17 |
| 33 | noted | 2.1 | 1.1 | 1.03 | bilaterally | 0.2 | 0.4 | -0.17 |
| 34 | involving | 1.2 | 0.2 | 1.01 | artifact | 0.1 | 0.3 | -0.16 |
| 35 | line | 1.1 | 0.1 | 0.99 | window | 0.2 | 0.3 | -0.16 |
| 36 | tegmen | 1.0 | 0.0 | 0.95 | dilatation | 0.0 | 0.2 | -0.15 |
| 37 | hematoma | 1.0 | 0.1 | 0.92 | neural | 0.0 | 0.2 | -0.15 |
| 38 | portion | 1.0 | 0.1 | 0.92 | otosclerosis | 0.1 | 0.2 | -0.14 |
| 39 | fragment | 0.9 | 0.0 | 0.89 | without | 1.7 | 1.8 | -0.14 |
| 40 | frontal | 1.0 | 0.1 | 0.85 | imaged | 0.1 | 0.3 | -0.14 |

## Appendix F: Evaluation of the Number Of Keywords and Performances on Five Different Models

Random Forest

| features | Accuracy | F1 | Precision | Recall |
|---|---|---|---|---|
| 2 | 0.793 | 0.618 | 0.687 | 0.625 |
| 3 | 0.951 | 0.913 | 0.93 | 0.91 |
| 4 | 0.946 | 0.899 | 0.93 | 0.89 |
| 5 | 0.94 | 0.881 | 0.93 | 0.87 |
| 10 | 0.964 | 0.926 | 0.983 | 0.89 |
| 20 | 0.976 | 0.95 | 1 | 0.915 |
| 30 | 0.97 | 0.935 | 1 | 0.89 |
| 50 | 0.969 | 0.93 | 1 | 0.885 |
| 100 | 0.97 | 0.935 | 1 | 0.89 |
| 200 | 0.964 | 0.921 | 1 | 0.865 |
| 300 | 0.964 | 0.916 | 1 | 0.865 |
| 500 | 0.964 | 0.916 | 1 | 0.865 |

Logistic Regression

| features # | Accuracy | F1 | Precision | Recall |
|---|---|---|---|---|
| 2 | 0.775 | 0.495 | 0.675 | 0.4 |
| 3 | 0.915 | 0.821 | 0.92 | 0.78 |
| 4 | 0.91 | 0.82 | 0.92 | 0.76 |
| 5 | 0.909 | 0.807 | 0.927 | 0.755 |
| 10 | 0.915 | 0.814 | 0.95 | 0.75 |
| 20 | 0.922 | 0.843 | 0.913 | 0.81 |
| 30 | 0.921 | 0.841 | 0.907 | 0.805 |
| 50 | 0.915 | 0.815 | 0.94 | 0.755 |
| 100 | 0.933 | 0.844 | 0.98 | 0.77 |
| 200 | 0.939 | 0.855 | 1 | 0.77 |
| 300 | 0.945 | 0.874 | 1 | 0.795 |
| 500 | 0.945 | 0.874 | 1 | 0.795 |

SVM

| features | Accuracy | F1 | Precision | Recall |
|---|---|---|---|---|
| 2 | 0.769 | 0.474 | 0.658 | 0.375 |
| 3 | 0.94 | 0.904 | 0.906 | 0.915 |
| 4 | 0.946 | 0.901 | 0.923 | 0.895 |
| 5 | 0.94 | 0.886 | 0.923 | 0.87 |
| 10 | 0.946 | 0.9 | 0.942 | 0.87 |
| 20 | 0.946 | 0.904 | 0.93 | 0.895 |
| 30 | 0.952 | 0.916 | 0.937 | 0.915 |
| 50 | 0.951 | 0.909 | 0.93 | 0.915 |
| 100 | 0.976 | 0.95 | 0.98 | 0.935 |
| 200 | 0.975 | 0.949 | 0.98 | 0.93 |
| 300 | 0.975 | 0.949 | 0.98 | 0.93 |
| 500 | 0.975 | 0.949 | 0.98 | 0.93 |

Xgboost

| Word | Accuracy | F1 | Precision | Recall |
|---|---|---|---|---|
| 2 | 0.793 | 0.618 | 0.687 | 0.625 |
| 3 | 0.94 | 0.902 | 0.913 | 0.91 |
| 4 | 0.94 | 0.902 | 0.913 | 0.91 |
| 5 | 0.928 | 0.87 | 0.913 | 0.87 |
| 10 | 0.928 | 0.879 | 0.897 | 0.89 |
| 20 | 0.964 | 0.937 | 0.947 | 0.94 |
| 30 | 0.958 | 0.924 | 0.94 | 0.92 |
| 50 | 0.958 | 0.924 | 0.94 | 0.92 |
| 100 | 0.964 | 0.924 | 0.89 | 0.89 |
| 200 | 0.964 | 0.937 | 0.947 | 0.94 |
| 300 | 0.958 | 0.922 | 0.947 | 0.915 |
| 500 | 0.958 | 0.922 | 0.947 | 0.915 |

Decision Tree

| features | Accuracy | F1 | Precision | Recall |
|---|---|---|---|---|
| 2 | 0.817 | 0.621 | 0.751 | 0.575 |
| 3 | 0.952 | 0.922 | 0.93 | 0.91 |
| 4 | 0.922 | 0.857 | 0.872 | 0.87 |
| 5 | 0.904 | 0.848 | 0.89 | 0.85 |
| 10 | 0.916 | 0.87 | 0.826 | 0.87 |
| 20 | 0.928 | 0.869 | 0.903 | 0.89 |
| 30 | 0.928 | 0.866 | 0.86 | 0.85 |
| 50 | 0.922 | 0.888 | 0.861 | 0.89 |
| 100 | 0.945 | 0.907 | 0.918 | 0.885 |
| 200 | 0.933 | 0.897 | 0.947 | 0.915 |
| 300 | 0.952 | 0.916 | 0.927 | 0.94 |
| 500 | 0.921 | 0.866 | 0.867 | 0.82 |

## Appendix G: Correspondence with Peer Reviewers for Study 2, Interpretable Machine Learning Text Classification for Clinical Computed Tomography Reports – A Case Study of Temporal Bone Fracture, in Computer Methods and Programs in Biomedicine Update:

We thank the reviewers for their careful reading of the manuscript and their constructive comments. We carefully considered all the comments and made significant revisions to improve and clarify the manuscript. Because we extensively revised the manuscript, we summarized the major changes in "Major changes of the revised manuscript" for your convenience of reading.

Reviewer 1: I had the pleasure to review the manuscript "Interpretable Machine Learning Text Classification for Computed Tomography reports - A Case Study of Temporal Bone Fracture." In this investigation, the authors demonstrated the application of Random Forest (RF), SVM, and decision trees classifier combined with commonly used NLP methods (BOW, TF-IDF) in classifying fracture cases from non-fracture cases. Word Frequency Score (WFS) and LIME were used for interpretability. A limitation of this paper, as previous reviewers have noted, is that the sample size was only 164 reports, which hinders the experimental conclusions. In addition, these 164 reports came from a single center with patients in a very small age range (60-65), which severely limits the clinical value of the proposed models. Despite its limitations, I think this manuscript is still a solid paper that illustrates a use case for how tools in explainable AI could be used to improve the transparency of ML models made for clinical purposes. The authors have revised and answered previous reviewers' comments appropriately. My main feedback includes the need for a baseline model, try boosting algorithms, and more explanation of the top predictors as well as rephrase some sentences in the discussion and conclusions.

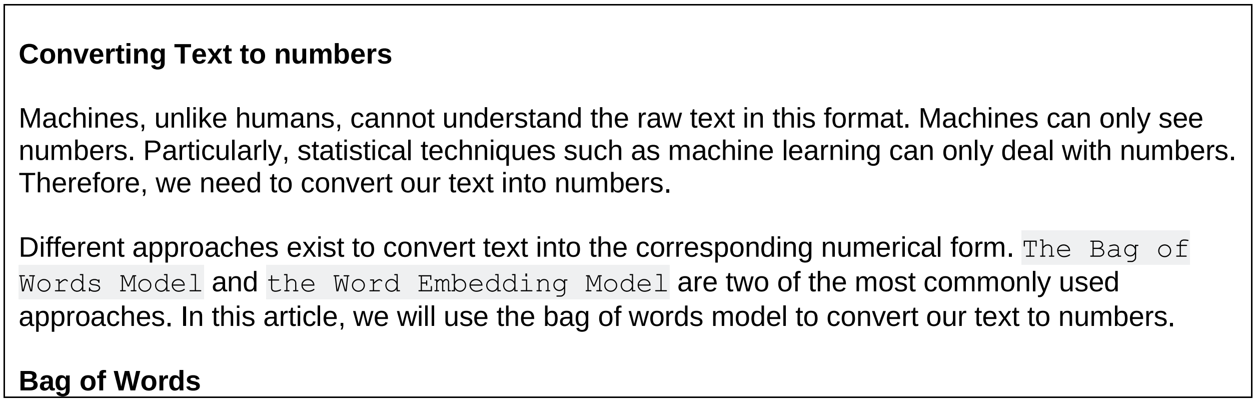
Response: Thank you for your kind comment. We believe our revised manuscript has improved descriptions of the baseline model, boosting algorithms, and more explanation of predictors. We also make our best effort to rephrase some language parts.

Major Comments

*1. In the 'Development of ML models' section, please explain why you set the minimum frequency limit threshold to 4 and the maximum frequency to 70%. Shouldn't these parameters be optimized via a parameter search?*

Response: These parameters could have been optimized via a parameter search. However, we use these numbers based on our observation of word frequency analysis and a full list of top frequency words. Please check the list in the supplemental files. From the list of words and the list of frequencies, we find that a minimum frequency of five and a maximum frequency of 70% is a threshold to produce a reliable performance. Searching for the best optimum parameter is possible, but the optimized parameters are not likely to change the interpretation results. Because our study focused more on interpreting results, we relied on reasonably good parameters to create a dataset.

Also, please see the pieces of code and explanations from the Jupyter Notebook. The notebook is available in the supplemental files.



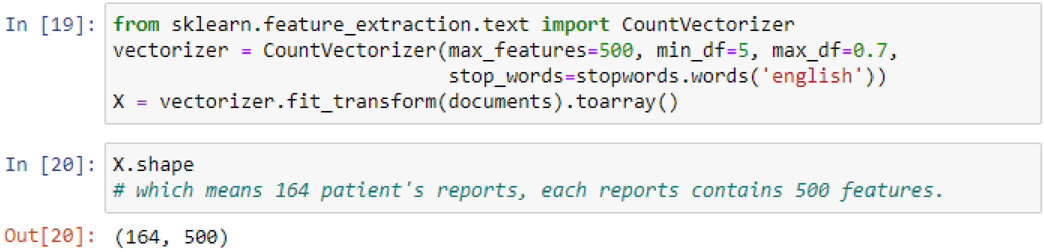
**Converting Text to numbers**

Machines, unlike humans, cannot understand the raw text in this format. Machines can only see numbers. Particularly, statistical techniques such as machine learning can only deal with numbers. Therefore, we need to convert our text into numbers.

Different approaches exist to convert text into the corresponding numerical form. `The Bag of Words Model` and `the Word Embedding Model` are two of the most commonly used approaches. In this article, we will use the bag of words model to convert our text to numbers.

**Bag of Words**

The following script uses the bag of words model to convert text documents into corresponding numerical features:



```
In [19]: from sklearn.feature_extraction.text import CountVectorizer
         vectorizer = CountVectorizer(max_features=500, min_df=5, max_df=0.7,
                                      stop_words=stopwords.words('english'))
         X = vectorizer.fit_transform(documents).toarray()
```

```
In [20]: X.shape
         # which means 164 patient's reports, each reports contains 500 features.

Out[20]: (164, 500)
```

The script above uses `CountVectorizer` class from the `sklearn.feature_extraction.text` library. There are some important parameters that are required to be passed to the constructor of the class. The first parameter is the `max_features parameter`, which is set to 1500. This is because when you convert words to numbers using the bag of words approach, all the unique words in all the documents are converted into features. All the documents can contain tens of thousands of unique words. But the words that have a very low frequency of occurrence are unusually not a good parameter for classifying documents. Therefore, we set the `max_features` parameter to 1500, which means that we want to use 500 most occurring words as features for training our classifier.

The next parameter is `min_df` and it has been set to 5. This corresponds to the minimum number of documents that should contain this feature. So we only include those words that occur in at least 5 documents. Similarly, for the `max_df`, feature the value is set to 0.7; in which the fraction corresponds to a percentage. Here 0.7 means that we should include only those words that occur in a maximum of 70% of all the documents. Words that occur in almost every document are usually not suitable for classification because they do not provide any unique information about the document.

Finally, we remove the `stop words` from our text since, in the case of this analysis, stop words may not contain any useful information. To remove the stop words we pass the `stopwords` object from the `nltk.corpus` library to the `stop_words` parameter.

The `fit_transform` function of the `CountVectorizer` class converts text documents into corresponding numeric features.

2. For the parameter search of the number of features, where did the feature ranking (from top most feature to the bottom feature) come from?

Response: The feature ranking comes from the frequency of the words that occurred in our entire clinical document set. In other words, we used the 500 most frequently occurring words as a feature to train the classifier. The most occurring words, of course, exclude any stop words listed in the natural language toolkit (NLTK), so these top words show semantics for positive and negative documents.

3. The explanation of LIME in the 'Interpretation of ML models' sounds vague. Please explain how the local model was constructed.

Response: We used the LIME package and included the text explainer. Our pipeline completely follows the pipeline on the documentation example. Please check the documentation for Text Explainer.

https://eli5.readthedocs.io/en/latest/tutorials/black-box-text-classifiers.html#textexplainer

To answer the question of how the local model was constructed, Text Explainer generated texts similar to the document (by removing and sampling words in the document), and then trained a white-box classifier which predicts the output of the black-box classifier. The explanation we saw is for this white-box classifier.

This approach follows the LIME algorithm; for text data, the algorithm is pretty straightforward:

1. generate slightly modified versions of the text.
2. predict probabilities for these modified texts using the black box classifier;
3. train another classifier (one of those eli5 supports) which tries to predict output of a black-box classifier on these texts.

The algorithm works because even though it could be hard or impossible to approximate a black-box classifier globally (for every possible text), approximating it in a small neighborhood near a given text often works well, even with simple white-box classifiers.

Because the local model is always an approximation of a black-box model, we included the mean_KL_divergence and accuracy score in the white-box classifier. It usually assigns the same labels as the black-box classifier on the dataset we generated, and its predicted probabilities are close to those predicted by our machine learning pipeline. The most valuable thing is that the local model could provide weights and word-level visualizations for the contribution of fracture classification.

As for the problem of 'Interpretation of ML models' sounds vague, we described the concept in a clearer way in the manuscript: LIME provides the word-level evaluation on how each word contributes to the classification results.

4. Regarding Fig 3: what is the significant of 'left', 'right', and 'canal' in terms of interpretability? An explanation for why these words is important and meaningful to a physician would be beneficial.

Response: Please see an example of documents, which shows the words "left", "right," "fracture," "canal".

Document #1:

1. Old comminuted fracture of the right middle cranial fossa with multiple bullet fragments lodged within it as described above. There is disruption/disolution of the right ossicular chain but the inner ear structures are intact.

2. Adjacent residual right mastoid air cells are chronically opacified.

Examination reviewed by Dr. [NAME] and reported findings confirmed by Dr. [NAME].

Clinical Indication: Post traumatic right otalgia.

Techniques: 0.625 mm thick contiguous axial scans of temporal bones were acquired. Coronal reformats were generated and reviewed.

Comparison: None.

Findings: Again visualized are severely comminuted old fractures of the right middle cranial fossa with multiple bullet fragments within the bones of the right skull base the right middle ear cavity and right anterior mastoid air cells. The roof of the right middle year cavity is dehiscent and there are dislocations/resorptions of components of the right ossicular chain. Only the body of the right incus is well visualized.

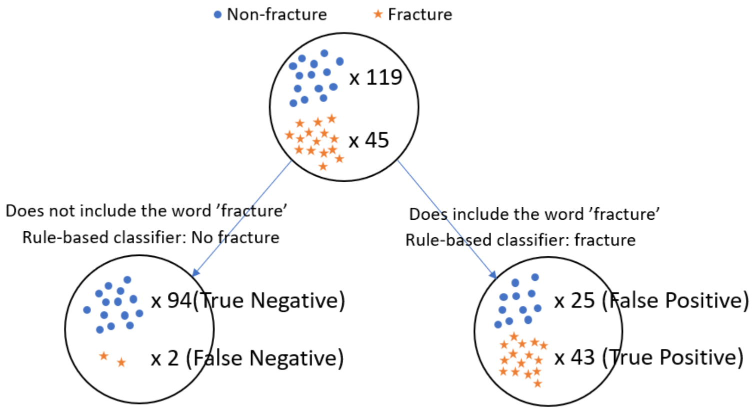
> Multiple bullet fragments are lodged within the clivus sphenoid bone prevertebral soft tissues and in the infratemporal fossa. The residual mastoid air cells are opacified.
>
> The left mastoid air cells appear well-aerated. The left middle ear cavity and ossicular chain are preserved. The left mastoid air cells appear unremarkable.
>
> Bilateral inner ear structures appear normal in morphology and density. Internal auditory canals appear symmetrical and normal in size bilaterally. Vestibular aqueducts are not dilated.

The terms "left" and "right" "canal" appear frequently in this document as the condition of parts of anatomical words. These anatomical concepts are often a major part of descriptions in temporal CT clinical documents. A hypothesis is that once a fracture is discovered, physicians tend to provide more detailed descriptions of the anatomical parts and tissues. In clinical settings, fracture diagnosis only consists of a small percentage of documents. Physicians may take more time on abnormal CT images, and draft more detailed documents, thoroughly describing all parts of the body system. This makes the "right" and "left" "canal" show up more often in fracture texts. For non-fracture documents, we hypothesize that physicians tend to use more summarized words, such as "All [parts] appear intact, in normal morphology and density." This description is more likely to be non-fracture documents, will be shorter, and will use fewer anatomical terms.

5. Regarding Fig 3: This figure suggests that the high precision probably comes from the fact that fracture reports typically contain words like 'fracture' or 'temporal.' Also, words like 'lung' or 'calcification' imply a non-fracture case because the CT images are of the lungs or the heart. This made me think perhaps a simple rule-based model for these specific words may suffice the classification task. Would be nice to include a model as such to compare with the existing models.

Thank you very much for your suggestion. We use the word "fracture" to build a simple one-rule classifier. We wanted to keep the rule simple, because it is common to conduct keyword searches and quickly determine the classifications. This would be a good baseline to reflect the actual scenarios for bone fracture classification.

In this case, the rule-based classifier would classify documents with "fracture" as positive and documents without "fracture" as negative. This would serve as a baseline model. In supplemental files, we included the rule-based classifier and used the "if" condition to construct the classifier on our documents. We then count the true positive, false positive, true negative, and false negative cases. We measure F1, precision, and recall. Please see the following figure:


● Non-fracture
★ Fracture
x 119
x 45
Does not include the word 'fracture'
Rule-based classifier: No fracture
Does include the word 'fracture'
Rule-based classifier: fracture
x 94(True Negative)
x 2 (False Negative)
x 25 (False Positive)
x 43 (True Positive)

When we adopt "fracture" as the only rule of the rule-based classifier:

Precision = TP/(TP+FP) = 43/(43+25) = 0.632

Recall = TP/(TP+FN) = 43/(43+2) = 0.956

F1 = 2*Precision*recall/(precision + recall) = 0.761

Also, we provided a supplemental file named "rule-based classifier" that provided case-by-case prediction results and rule-based model details.

From the rule-based model's result, the F1-score is 0.761, which is significantly lower than our machine learning-based classifier. TP, FP, FN, and TN cases, we see more false positive cases than false negative cases. This indicates even a simple rule-based classifier will not be likely to miss a fracture diagnosis. The precision is a bit lower than our machine learning model. To increase the precision, therefore, it makes sense to build more complicated rules or to adopt the most frequent words as machine learning features.

As we investigated in related work, the development of rule-based classifiers was mostly in the 1990s. It must be admitted that a simple rule-based classifier cannot handle complex clinical text classification tasks. If we apply multiple rules, the performance may improve, but interpretation becomes a problem again. The rule-based classifier may not adapt to the ever-changing word usage in medical documents. The machine learning models, however, can overcome this problem. Therefore, building more complicated rules is no longer a focus of our study. We may not use a 20-year-old rule-based classifier as a baseline. Instead, starting from the machine learning model would be a better choice. Based on the rule-based classifier's performance, the ML model's development convenience, and comparisons, we decided to not include the rule-based classifier as a baseline model.

*6. Regarding Fig 4: The comparison between RF, SVM, Decision Tree, and Logistic Regression provides little values. Have you tried boosting methods (XGBoost, LightGBM, CatBoost)? You could make the text sizes in the axes and legends bigger.*

Thank you for the suggestion! We included an extra XGBoost model and provided performance values in two figures. Generally speaking, the XGBoost model shows comparable performance to the Random Forest model. The precision, recall, and F1 values are similar. The XGBoost model's performance has been added to the revised figures.

*7. In addition, you mentioned that from these results that you chose 500 as the number of topics, but the RF plot suggests that 20 is the best number.*

Response: Thank you for pointing out this issue. Please see the graph below to see the relationship between the machine learning model, the number of features in the dataset, and the Evaluation Performances. This figure is a simplified visualization of previous figures. The performance metrics did not change.

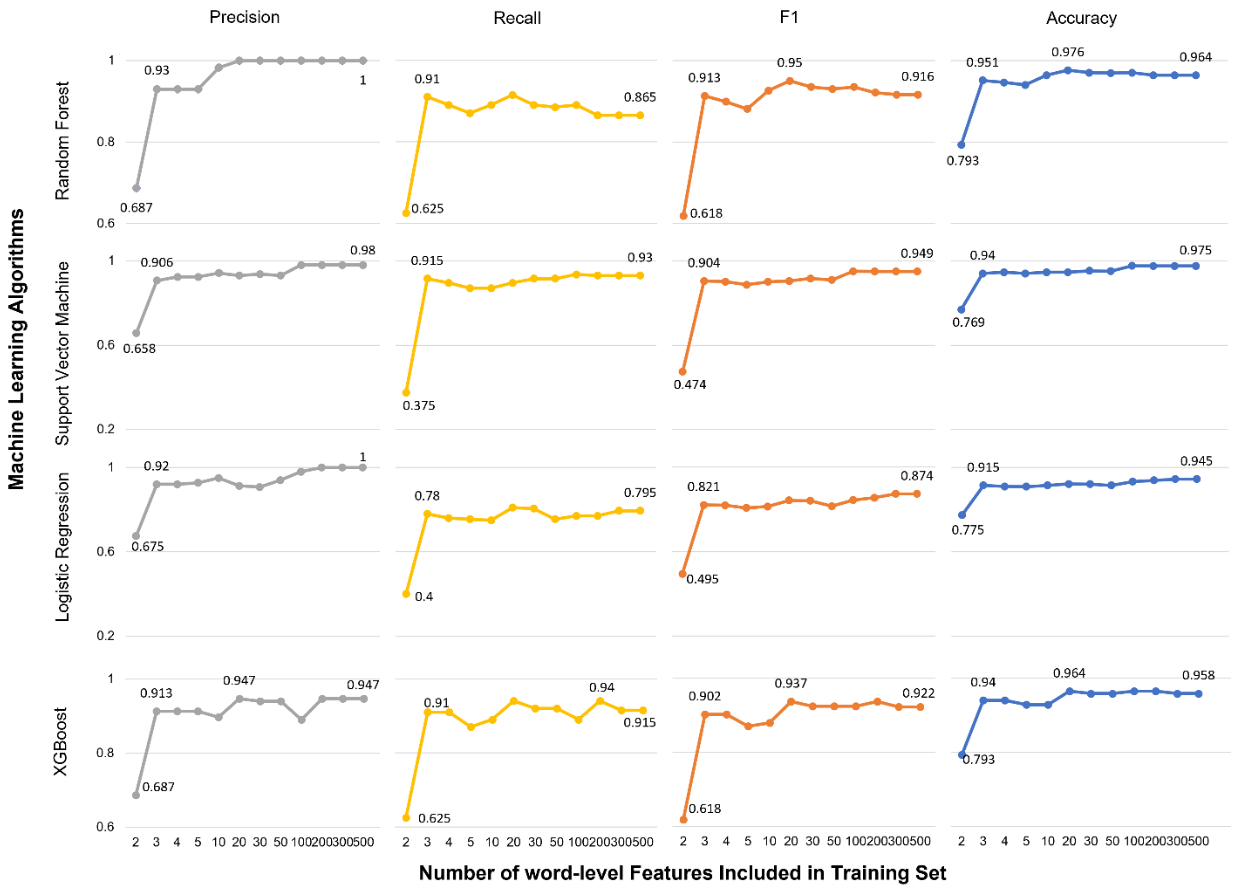


Figure 4: Relationships between classification model’s performance, number of selected features, and evaluation performances for Random Forest, XGboost, Support Vector Machine, and Logistic Regression model.

In this set of figures, we can quickly find that most performance is better as the number of selected words gets larger. Therefore, we keep our main models to a 500-word limit.

*8. I think either Fig 4 or Fig 5 should be in the main body, and the other one could be moved to the Supplement.*

Response: We moved figure 5 into the supplemental files section. Figure 3 would show the model's performance from a comprehensive perspective.

*9. Regarding Fig 6: The AUC is very high, that made me think whether the problem of classifying fracture/non-fracture cases is challenging enough to utilize ML. If it is a relatively simple problem, a simple model would suffice. This could be answered by including a simple rule-based model in the analysis to serve as a baseline.*

We built a rule-based classifier and reported the performance. Please refer to the response in comment 5.

Here is the answer to the question of whether the problem of classifying fracture and non-fracture cases is challenging enough to utilize ML: Machine learning models need to be used because they work much better than a simple rule-based classifier.

*10. Figure 7 and 8 could be merged into one. Similarly, some figures are redundant (see below), and 10 total figures is a lot. Please merge some figures and/or put some in the Supplements.*

*11. Fig 10 is nice. Not a new method but a nice and definitely interpretable visualization to summarize the logic behind a model.*

Comment 10 and 11 are both related to figures, so we combined them and responded to the comments together:

After we discussed it with the coauthors, we agreed to make the following edits to the charts:

| Action | Which Chart(s) |
|---|---|
| Leave these figures on the main text: | • Overview of our study<br>• Comparison of gaps between fracture and non-fracture reports |

| | |
|---|---|
| | • Relationships between classification model's performance, number of selected features, and evaluation performances<br>• Merge two figures: Text Explainer's Evaluation (original Figure 7) and Top Features and its weight by Text explainer (original Figure 8)<br>• A visualization of decision trees. |
| Delete this chart: | • Overview of Text Explainer, how text explainer explains clinical CT documents |
| Move these to supplemental files: | • Accuracy score and the Kullback-Leibler divergence score to Evaluate the reliability of LIME Interpretation<br>• AUROC Figure of random forest model<br>• Classification Performance for Support Vector Machine, Logistic Regression and Random Forest Models |

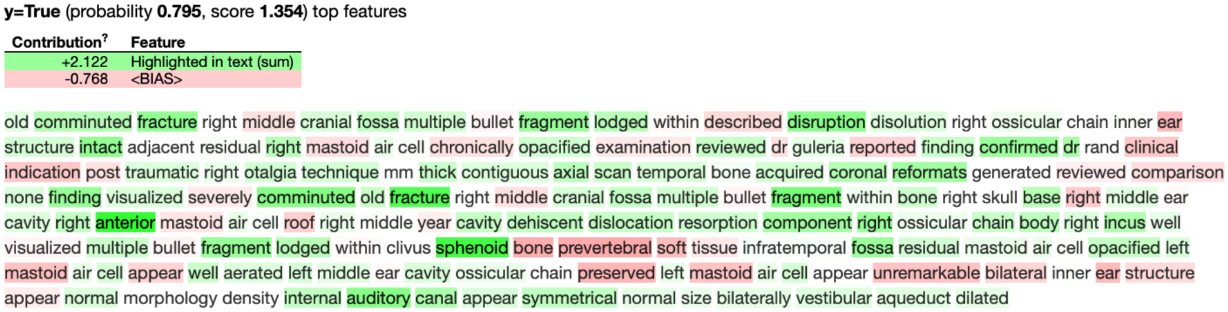
Please see our updated manuscript for detailed changes of the charts.

*12. I appreciate the author's answer to a reviewer's comment regarding the reason that authors used BOW instead of Word-2-vec. Would be nice to include it in the main text.*

Response: Thank you for the suggestion. We moved the justification of using BOW instead of word-2-vec into the main text.

*13. From my understanding, LIME can only explain individual reports, and not globally. Since LIME trains a local linear model around the individual prediction to approximate the prediction of the 'black box' model. Thus, I am confused as whether the model could get an aggregated explanation?*

Here is the original figure 7 and figure 8:

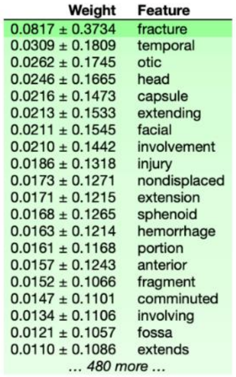


**y=True** (probability **0.795**, score **1.354**) top features

| Contribution? | Feature |
|---|---|
| +2.122 | Highlighted in text (sum) |
| -0.768 | <BIAS> |

old comminuted fracture right middle cranial fossa multiple bullet fragment lodged within described disruption disolution right ossicular chain inner ear structure intact adjacent residual right mastoid air cell chronically opacified examination reviewed dr guleria reported finding confirmed dr rand clinical indication post traumatic right otalgia technique mm thick contiguous axial scan temporal bone acquired coronal reformats generated reviewed comparison none finding visualized severely comminuted old fracture right middle cranial fossa multiple bullet fragment within bone right skull base right middle ear cavity right anterior mastoid air cell roof right middle year cavity dehiscent dislocation resorption component right ossicular chain body right incus well visualized multiple bullet fragment lodged within clivus sphenoid bone prevertebral soft tissue infratemporal fossa residual mastoid air cell opacified left mastoid air cell appear well aerated left middle ear cavity ossicular chain preserved left mastoid air cell appear unremarkable bilateral inner ear structure appear normal morphology density internal auditory canal appear symmetrical normal size bilaterally vestibular aqueduct dilated

*Figure 7: Text Explainer's Evaluation on word's Contribution using Random Forest Algorithm*



| Weight | Feature |
|---|---|
| 0.0817 ± 0.3734 | fracture |
| 0.0309 ± 0.1809 | temporal |
| 0.0262 ± 0.1745 | otic |
| 0.0246 ± 0.1665 | head |
| 0.0216 ± 0.1473 | capsule |
| 0.0213 ± 0.1533 | extending |
| 0.0211 ± 0.1545 | facial |
| 0.0210 ± 0.1442 | involvement |
| 0.0186 ± 0.1318 | injury |
| 0.0173 ± 0.1271 | nondisplaced |
| 0.0171 ± 0.1215 | extension |
| 0.0168 ± 0.1265 | sphenoid |
| 0.0163 ± 0.1214 | hemorrhage |
| 0.0161 ± 0.1168 | portion |
| 0.0157 ± 0.1243 | anterior |
| 0.0152 ± 0.1066 | fragment |
| 0.0147 ± 0.1101 | comminuted |
| 0.0134 ± 0.1106 | involving |
| 0.0121 ± 0.1057 | fossa |
| 0.0110 ± 0.1086 | extends |
| *… 480 more …* | |

*Figure 8: Top Features and its weight by Text explainer using Random Forest Algorithm*

First, we would like to clarify that figures 7 and 8 are not relevant. We use the word "contribution" in figure 7, and "weight" in figure 8. We recognize that putting Figure 7 and Figure 8 together may cause confusion very easily. Here are explanations for figures 7 and 8:

Figure 7 is a visualization of a text explainer. The text explainer's only job is to evaluate each word's contribution to each document. The text explainer's evaluation is based on an approximated white-box model, as we mentioned. In the explainer, each word's contribution could be positive (shown in green) or negative (shown in red). The red word means a contribution to negative classifications. The green word means a contribution to positive classification. For example, in Figure 7 text example, we can say the words "comminuted" and "fracture" contribute to a positive result. "Unremarkable" contributes to negative results. In other words, in the text explainer, the word has two directions. Therefore, figure 7’s contribution is only related to each document. Figure 7’s word-level contributions are not aggregated evaluations.

In Figure 8, LIME evaluated the weight of the completed model based on a global view. The weight is based on the entire document set and does not relate to any form of classification results. We can say that "weight" is an alternative expression of "feature importance". Of course, they are based on different algorithms, but both key concepts are to evaluate which feature serves as a deciding factor in classification.

The "fracture" has a weight of 0.0817. In Figure 8, the information only shows "fracture" is often used as the primary deciding factor in decision trees from the random forest. "Fracture" serves as a key deciding criteria in many decision trees. Therefore, the weight is calculated based on the completed machine learning model, which is trained based on the entire training set. Therefore, the figure 8 weight list is based on aggregated evaluations.

We thank you for your comments. With your comment, we find that our paragraph may cause confusion in the discussion section "Interpretation of Machine Learning Models." To avoid confusion, we provide the following additional clarification in the main text:

Figure 7 shows a visualization of a text explainer. In this text explainer, each word has been assigned a contribution score, showing the words lead to positive or negative classification. The text explainer’s evaluation is based on the individual text level.

The feature list in Fig. 8 displays the random forest model's most important word list. The word list is aggregated from multiple decision trees. The selection of each word is calculated by whether the word serves as a deciding factor in a decision tree. Higher weight words are often used as a key factor in the classification results. The feature list corresponds to the text explainer's assessment of figure 7.

*14. Regarding the second paragraph of the 'Summary of Text feature analysis' in the Discussion comment: I don't quite agree with the remark on using highly specialized language patterns are highly conserved. Sometimes it is the case that clinical texts are unavoidably full of domain-specific keywords, and use of a specialized corpus instead of a general one could achieve higher performance.*

Response: We agree with your argument. Sometimes, clinical texts are unavoidably full of domain-specific keywords.

Use of a specialized corpus instead of a general one could achieve higher performance. We included this direction as a part of our future work. Our work’s next step is to only visualize the medical specialized words in clinical documents. This can reduce the effect of other irrelevant words and make a more accurate prediction.

Regarding our statements in the last paragraph of the 'Summary of Text feature analysis' in the Discussion, we have modified the texts, and added the following statements:

Physicians often draft reports with highly specialized medical terms. These medical terms often serve as reliable predictors. According to our results, we suggest investigating if non-experts can easily interpret the medical terms.

Our LIME results show that specialized medical words contribute significantly to classification.  As a result, we believe that interpretable AI has the potential to help explain more complex diagnostic conditions. This pipeline can also be used to interpret other clinical texts, classify diagnoses, and give an explanation that is easy to understand.

*15. I don't quite agree with the claim in the Discussion that LIME is significantly easier to interpret than decision tree, and many would say the same. This claim is too strong in my opinion. Each has its own merits. Perhaps LIME could be slightly more suitable for non-numeric data.*

Response: Thank you for your suggestion. However, we believe our claim is reasonable, especially for text data. We did not change our claim but have modified the claim slightly. Our scope of discussion focused on text data.

We believe LIME would be significantly more explainable than a decision tree. Our decision tree example only includes five keywords. However, a real decision tree has significantly more depth and more nodes. A short decision tree like this example is easy to understand, but in most clinical cases, the tree must be very deep to achieve satisfactory performance. The actual decision tree often reaches 100 or more nodes. In this case, the physician must iterate through ten levels of depth and choose from ten different rules. This is not practical. If a similar LIME network was constructed, the visualization would only show a list of weights like in figure 8. This list makes it easier for physicians to determine the importance of each word to the classification results.

*16. Regarding Table 1: not quite sure what it means for inherently transparent models to require deep AI knowledge. Please elaborate or word differently. In addition, there are several papers that argue otherwise (pro for fully transparent models and against post-hoc methods). This is a highly controversial topic that is also not quite related to the sole purpose of interpretable classification of fracture reports from texts.*

We agree that it is a highly controversial topic about choosing which forms of models to use. After carefully considering your comments, we believe deleting this part of the discussion would help focus on our topic of "Interpretable models on clinical document classification". As you mentioned, the topic of selecting which types of models is outside the scope of this study. Therefore, we decided to delete that part of the discussion. The deletion allows the paper to focus on the LIME and the WFS analysis parts of the work.

*17. Regarding the conclusions: conclusion #2 about RF achieved the highest accuracy, in my opinion, is not novel nor significant.*

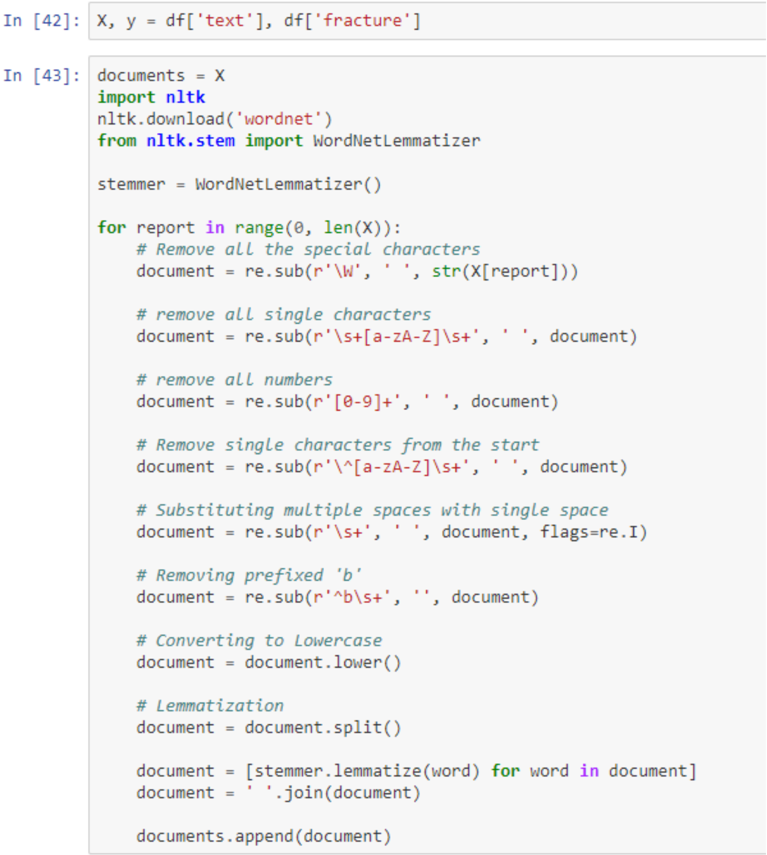
Response: Thank you for the comments. We deleted the RF part and focused more on the significance of our model's interpretation in the conclusion.

Minor Comments

*18. Regarding Text pre-processing: please provide in the Methods more details on which parts in the preprocessing step were automated and which packages were used for them. It will only take one or two sentences.*

Response: We added a brief description of the text processing on the section about the following text pre-processing steps.

We removed all non-word elements in clinical reports, including numbers, punctuation, and special characters. We converted all words to lowercase letters. We perform these changes using regular expressions.  Then, we followed a Natural Language Toolkit stop-word list to remove stop words, lemmatized words, and incorrect spellings and acronyms. All words were free of noun declination and verb conjugations. The supplemental code book shows how we pre-process the documents.



```
In [42]: X, y = df['text'], df['fracture']
```

```
In [43]: documents = X
         import nltk
         nltk.download('wordnet')
         from nltk.stem import WordNetLemmatizer

         stemmer = WordNetLemmatizer()

         for report in range(0, len(X)):
             # Remove all the special characters
             document = re.sub(r'\W', ' ', str(X[report]))

             # remove all single characters
             document = re.sub(r'\s+[a-zA-Z]\s+', ' ', document)

             # remove all numbers
             document = re.sub(r'[0-9]+', ' ', document)

             # Remove single characters from the start
             document = re.sub(r'\^[a-zA-Z]\s+', ' ', document)

             # Substituting multiple spaces with single space
             document = re.sub(r'\s+', ' ', document, flags=re.I)

             # Removing prefixed 'b'
             document = re.sub(r'^b\s+', '', document)

             # Converting to Lowercase
             document = document.lower()

             # Lemmatization
             document = document.split()

             document = [stemmer.lemmatize(word) for word in document]
             document = ' '.join(document)

             documents.append(document)
```

We thank you very much for your thoughtful review and valuable comments.

*Reviewer 2: The paper uses four machine learning techniques, Support Vector Machine(SVM), Decision Tree(DT), Logistic regression, and Random Forest, to do binary classification on a case study of Temporal Bone Fracture and classify 164 Electronic health reports into fracture cases and non-fracture cases.*

*Although the dataset size seems small, I have seen great articles for different diseases, such as "Clinical text classification of Alzheimer's drugs' mechanism of action," that perform the same task on a small dataset. Having small datasets for clinical text classification is common, and the way that the paper uses Machine Learning techniques instead of Neural Network-based (NN) models or transformers makes sense to me because using NN models and fine-tuning transformers needs large datasets.*

*However, the final goal of the paper is interpreting the models not achieving higher accuracy. So that is excellent work.*

Response: Thank you for your comments! We would like to discuss a few points:

1. We analyzed word frequency and the gaps between negative and positive documents. We supplied the analysis in supplemental files. In the analysis, there are fewer than 1000 valid words as features. For this size, we think neural network-based models or transformers would not achieve an advantage over machine learning models. Should we include a larger set of clinical documents and a more complex classification task, the neural network might achieve a significantly better outcome. In future work, we will consider neural networks and transformers and apply them to a larger set of clinical documents to process complex classification tasks.

2. We did not notice the paper "*Clinical text classification of Alzheimer's drugs' mechanism of action*" in the literature review. They are very valuable work. In this revision, we added the relevant publication and our comments in related work section.

Major Comments:

*1. If the paper's work is on interpretability, which I believe it is, I strongly recommend that the authors remove the Decision Tree (DT) from the models because, in essence, it is not a black-box model. It is a white box model. Moreover, One category of Stanford's interpretability proposed methods is trying to draw a DT model and write an Interpretable decision set (IDS), so considering the DT model as a block-box model is entirely wrong.*

Response: Thank you for pointing out the issue. We have removed all content related to decision trees from the models and performance evaluations. However, we still like to discuss the model interpretability in decision trees and use decision trees as an example of a white box model. because we focus on the model's interpretability. It is essential to mention that the decision tree is unique because it is an intrinsically interpretable machine learning model. We would like to compare this model with other explanation techniques, such as LIME.

*2. On the other hand, Both RF and SVM models are black box, so I believe your job on these two models makes sense.*

Response: Thank you for your comments.

*3. You need to talk about the dataset details. It is the only comment you have not addressed correctly. I know you provided the code, but I assume that the results are unreliable if the dataset is imbalanced in small datasets. Therefore, the interpretation is not correct. Please add a table and provide information about the train, test, and evaluation size.*

Response: First, we apologize for not using the most precise words in the manuscript. The 10-fold cross validation should be stratified by the 10-fold cross validation. The procedure was shown in the Jupyter notebook.

We acknowledge that we did not provide sufficient information on training, testing, and validation in the last revision. We are sorry if there is any confusion in the manuscript, but we believe our corrected language has reflected the actual case after revision. We added new descriptions of the cross validation and train/test split percentages so our study can be reproduced. For the concerns of imbalanced data and small data sets, we believe our experiment and manuscript provided the correct interpretation to the maximum extent possible. If the manuscript needs more description, please inform us what part of the model creation information is critical to the reader. If you have concerns about codes, please suggest modifications to our codes. We will perform additional experiments.

In this study, we adopted stratified k-fold cross-validation to provide a reliable model performance. It is a technique to stratify the sampling by the class label, and this technique can tackle small and imbalanced datasets.

For more details of stratified k-fold, please see the response to Comment 6.

*4. Please define the acronym for the first time the sequence is mentioned in the text and avoid using the whole sequence or redefining it in the rest of the text. For example, once you write Support Vector Machine (SVM), do not redefine the acronym again; for the rest of the article, use SVM.*

Response: Response: Thank you. We have corrected all acronym issues in the manuscript. We removed a few infrequently used acronyms. The removed words are electronic health record

(EHR), natural language processing (NLP), and bag of words (BOW). We revised them accordingly for the rest of the words..

*5. In the Limitation and future work section, you divide your future work into two aspects. You say the first one but never clearly mention the second one. I believe the paper's English needs to be checked by the writing center. Some sentences are not clearly described.*

Response: We have re-stated the two aspects of our future work, as follows:

First, we used labeled data in this preliminary study. While unlabeled data cannot be used for classification, it has the potential for unsupervised learning. We believe that by building an appropriate unsupervised model, it is possible to cluster CT reports into two categories based on text reports. Second, building a medically specialized text interpreter would highlight medical words only and achieve a clearer interpretation. For example, by adopting SNOMED-CT standards [52], it is possible to create a medical text interpreter. The model could limit the number of words to only medical terms. The word-level optimization may achieve better prediction and better interpretation.

For the language issues, we have extensively looked for English editing and proofreading services, and we have revised the manuscript to the best of our effort.

Minor Comments

*6. In the comments, you mentioned that the (45 patients + 119 patients) if these are the number of samples for fracture and non-fracture cases. Is the data imbalanced? How did you deal with this problem? How are you using 10-fold for a class with 45 samples?*

Firstly, we acknowledge we did not choose the most precise words for the manuscript. The 10-fold cross validation should be stratified 10-fold cross validation. The procedure was shown in the Jupyter notebook.

The data is imbalanced. In this study, we used 164 clinical texts, with 45 positive cases and 119 negative cases. The small data set is common for clinical reports, and other studies use similar small data set. The study includes the one you mentioned "Clinical text classification of Alzheimer's drugs' mechanism of action".  Many studies have adopted stratified k-fold to deal with the data imbalance problem. We used the same stratified 10-fold to avoid bias as much as possible.

A stratified K-fold is a cross-validator that divides the dataset into k-folds. Stratified is to ensure that each fold of a dataset has the same proportion of observations with a given label. Here is a figure that shows how stratified k-fold works:

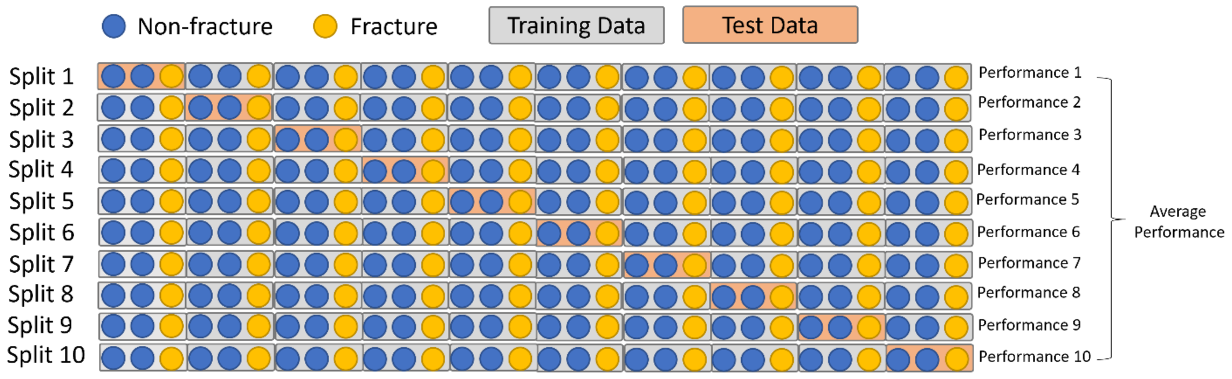


Figure 1. An example visualization of how stratified 10-fold splits the training set and test set

In this figure, fracture and non-fracture clinical texts are marked as yellow and blue dots, respectively. For easy plotting, we plotted 30 dots (20 negative, 10 positive), which is our entire data set.  The stratified 10-fold cross-validation method divides the training dataset into 10 folds. The first 9 folds are used to train a model, and the 10th fold serves as the test set. This process is repeated until each fold has a chance to be used as the holdout test set. A total of 10 models are fit and evaluated, and the model's performance is calculated as the mean of these runs.

Because stratified k-fold will evenly distribute the proportions of positive and negative cases and maintain a stable balance between test data and training data, It is widely accepted that stratified k-fold ensures that the proportion of positive to negative examples found in the original distribution is respected in all the folds. This is the best way to show an unbiased model's performance under a small data set.

As Dr Jason Brownlee writes in his article *"How to Fix k-Fold Cross-Validation for Imbalanced Classification":*

> It is a challenging problem as both the training dataset used to fit the model and the test set used to evaluate it must be sufficiently large and representative of the underlying problem so that the resulting estimate of model performance is not too optimistic or pessimistic.
>
> The two most common approaches used for model evaluation are the train/test split and the k-fold cross-validation procedure. Both approaches can be very effective in general, although train/test split can result in misleading results and potentially fail when used on classification problems with a severe class imbalance. Instead, the techniques must be modified to stratify the sampling by the class label, called stratified train-test split or stratified k-fold cross-validation.
>
> As we rarely have enough data to get an unbiased estimate of performance using a train/test split evaluation of a model. Using a stratified k-fold cross validation procedure would introduce minimal bias under limited dataset. The procedure has been shown to give a less optimistic estimate of model performance on small training datasets than a single train/test split. A value of k=10 has been shown to be effective across a wide range of dataset sizes and model types.

Therefore, we hope our explanation and figures can address your concerns about the small and imbalanced dataset. In our experiment, we considered the balance and small data set and chose the appropriate methods to avoid potential problems.

Again, we thank reviewers for the careful reading of the manuscript and constructive comments. We hope our revised paper has addressed all concerns by the reviewers in sufficient detail.