Risk Prediction and Stroke Recurrence Using Machine Learning and Healthcare Analytics: A Survey

www.surveyx.cn

Abstract

This survey paper examines the transformative role of machine learning (ML) and artificial intelligence (AI) in predicting stroke recurrence, emphasizing the potential to enhance clinical decision-making and patient outcomes. It highlights the significance of developing risk prediction models, particularly in the context of stroke recurrence, where timely interventions can alter patient trajectories. The paper is structured to provide a comprehensive overview, beginning with the significance of risk prediction and stroke recurrence, followed by an exploration of the statistical landscape and key risk factors. The survey delves into various ML techniques, including supervised and unsupervised learning, deep learning approaches, and ensemble and hybrid methods, showcasing their potential to improve predictive accuracy and model performance. Additionally, the paper addresses the challenges of data quality, integration, model complexity, interpretability, scalability, and bias, emphasizing the need for innovative frameworks and algorithms to enhance predictive models' reliability and applicability. Real-world case studies and applications demonstrate the practical implementation of these methodologies in clinical settings, highlighting their potential to transform stroke risk prediction and improve patient outcomes. The conclusion underscores the importance of integrating diverse data sources, ensuring model transparency and fairness, and prioritizing patient engagement in developing robust and trustworthy predictive models for stroke risk prediction.

1 Introduction

1.1 Significance of Risk Prediction and Stroke Recurrence

Accurate prediction of stroke recurrence is vital for improving clinical decision-making and patient outcomes, mirroring successful applications of machine learning (ML) and deep learning (DL) in other healthcare domains [1]. Developing robust risk prediction models refines clinical strategies, particularly for stroke recurrence, where timely interventions can significantly alter patient trajectories [2]. ML models have demonstrated superior predictive performance compared to traditional methods, such as early warning scores for sepsis prediction [3], and can similarly enhance stroke recurrence predictions, thereby improving healthcare strategies and patient management.

The complexities of intersecting group identities necessitate precise predictive models to address diverse patient needs and enhance health outcomes [4]. Accurate predictions are critical for managing hospital readmissions, as evidenced in conditions like sickle-cell disease, which underscores their relevance in stroke recurrence scenarios [5]. In resource-constrained settings, technologies such as photoplethysmography (PPG) can aid in developing deep learning-based cardiovascular disease risk scores, highlighting the importance of precise cardiovascular risk predictions, including stroke, to improve patient care [6].

Utilizing electronic health records (EHRs) for disease risk prediction enhances healthcare strategies by providing structured data for accurate assessments [7]. However, the increasing reliance on

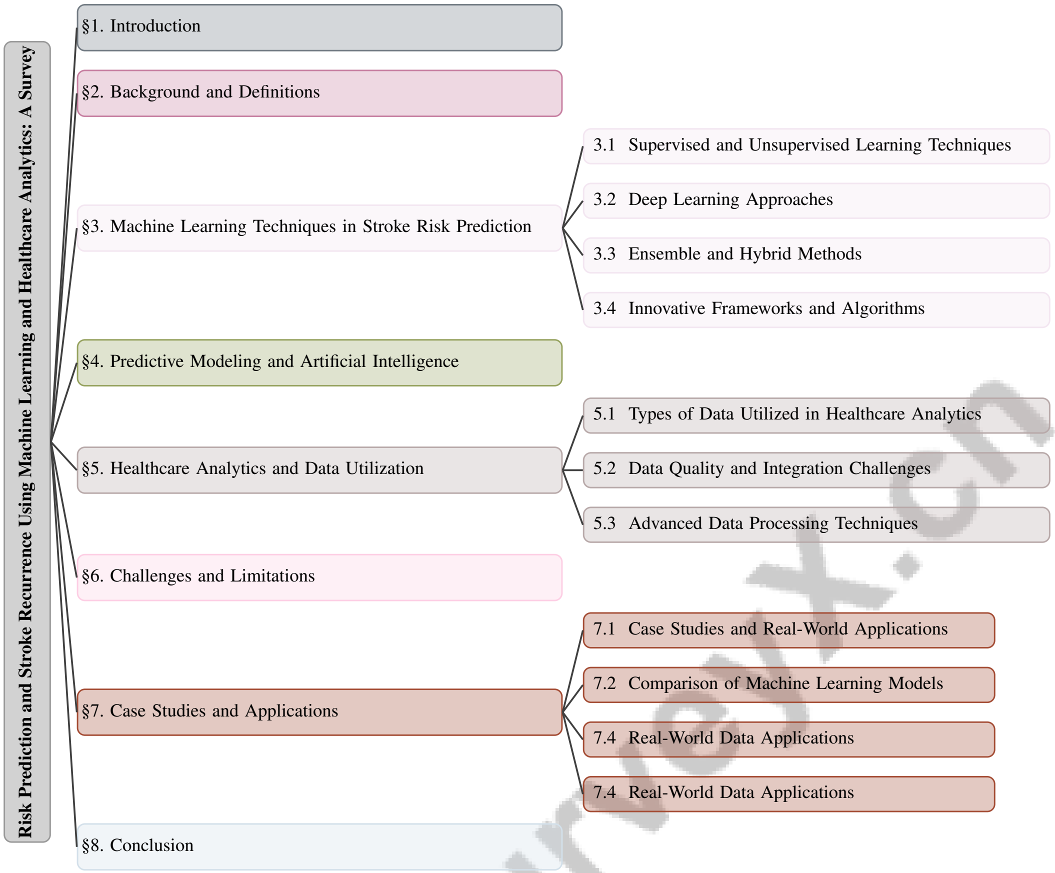

Risk Prediction and Stroke Recurrence Using Machine Learning and Healthcare Analytics: A Survey
§1. Introduction
§2. Background and Definitions
§3. Machine Learning Techniques in Stroke Risk Prediction
3.1 Supervised and Unsupervised Learning Techniques
3.2 Deep Learning Approaches
3.3 Ensemble and Hybrid Methods
3.4 Innovative Frameworks and Algorithms
§4. Predictive Modeling and Artificial Intelligence
§5. Healthcare Analytics and Data Utilization
5.1 Types of Data Utilized in Healthcare Analytics
5.2 Data Quality and Integration Challenges
5.3 Advanced Data Processing Techniques
§6. Challenges and Limitations
§7. Case Studies and Applications
7.1 Case Studies and Real-World Applications
7.2 Comparison of Machine Learning Models
7.4 Real-World Data Applications
7.4 Real-World Data Applications
§8. Conclusion

Figure 1: chapter structure

automated decision systems poses risks to population health equity, as these systems may perpetuate existing disparities, necessitating careful model development [8]. Identifying individuals at risk of cardiovascular diseases is fundamental to preventative cardiology, with accurate stroke recurrence predictions being integral to this effort [9].

Integrating multimodal healthcare data into disease risk prediction models can significantly improve patient outcomes, emphasizing the need for comprehensive approaches in stroke recurrence prediction [10]. These insights highlight the importance of developing sophisticated predictive models to tackle the multifaceted challenges of stroke recurrence, ultimately enhancing healthcare delivery and patient outcomes.

1.2 Structure of the Survey

This survey is organized into several key sections designed to systematically explore the complexities of stroke risk prediction and recurrence through machine learning and healthcare analytics. It begins with an introduction that emphasizes the significance of accurate risk prediction in stroke recurrence, illustrating its potential to transform healthcare strategies and improve patient outcomes. The background and definitions section follows, providing a comprehensive overview of stroke and its recurrence, elucidating the statistical landscape and key risk factors, along with precise definitions of terms such as machine learning, predictive modeling, artificial intelligence, and healthcare analytics, thus laying the groundwork for subsequent discussions.

The next section examines machine learning techniques in stroke risk prediction, covering a variety of algorithms and models, including supervised and unsupervised learning, deep learning approaches, and ensemble and hybrid methods. This is succeeded by a discussion of innovative frameworks and algorithms developed to enhance predictive accuracy, drawing insights from recent advancements in the field [11].

AI-generated, for reference only.

Subsequently, the role of predictive modeling and artificial intelligence is explored, focusing on the development and validation of predictive models for clinical use, their integration into healthcare practice, and the critical need for model interpretability and trustworthiness. The paper also addresses effective incorporation of these models into clinical settings to forecast stroke recurrence and improve patient management [12].

Healthcare analytics and data utilization are examined next, emphasizing the types of data used, such as electronic health records (EHRs), and the challenges related to data quality and integration. This section also discusses advanced data processing techniques essential for extracting valuable insights from complex healthcare datasets.

The survey identifies challenges and limitations associated with employing machine learning and AI for stroke risk prediction, addressing ethical and privacy concerns, model complexity, scalability, and the need to mitigate bias and ensure model generalizability [8]. These considerations are crucial for developing equitable and effective predictive models in healthcare.

Finally, the paper presents case studies and applications, showcasing real-world examples of successful machine learning and AI implementations in stroke risk prediction. This section compares various machine learning models and discusses the use of real-world data in developing predictive models, offering practical insights into the application of these technologies in clinical practice. The conclusion synthesizes key findings and suggests future research directions in this rapidly evolving field.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Overview of Stroke and Recurrence

Stroke is a significant global health challenge with high recurrence rates, necessitating effective predictive and management strategies. Cardiovascular diseases (CVDs) substantially contribute to morbidity and mortality, with stroke being a primary outcome [9]. Recurrence statistics, indicating a 7.4

Beyond conventional risk factors such as hypertension, diabetes, and hyperlipidemia, atrial fibrillation notably elevates stroke risk, necessitating robust prediction mechanisms [12]. Class imbalance in risk prediction models poses significant challenges, potentially skewing risk estimates if not properly addressed [2]. Utilizing diverse datasets, such as the UK Biobank, which includes 423,604 participants without baseline CVD, enhances our understanding of stroke and its recurrence [9]. The multimodal and heterogeneous nature of healthcare data necessitates comprehensive approaches for effective stroke recurrence prediction [10]. Research highlights AI's potential to enhance health outcomes and operational efficiencies, particularly in disease detection and risk prediction [8].

2.2 Key Terms and Definitions

Machine learning (ML), a branch of artificial intelligence, develops algorithms to learn from data and make predictions, facilitating the analysis of complex datasets like EHRs to predict disease risks and clinical events. Techniques such as random forest regression are valued for incorporating numerous predictors and modeling complex interactions nonparametrically [13]. These methods are vital for clinical prediction models (CPMs), traditionally predicting single outcomes but expandable to complex frameworks like probabilistic classifier chains and Bayesian probit models [14].

Predictive modeling uses statistical techniques to forecast future outcomes based on historical data, essential for stroke risk prediction. Traditional models, often limited in predictors, show suboptimal performance across diverse patient groups [9]. Advanced modeling benefits from integrating multimodal healthcare data, offering comprehensive disease risk insights [10]. Combining survival analysis with binary outcome prediction exemplifies a holistic evaluation of model performance across statistical methodologies [5].

Artificial intelligence (AI) in healthcare simulates human cognitive functions, enhancing decision-making by automating complex processes and integrating diverse data sources for improved patient outcomes [15]. AI's role is crucial in developing equitable models that perform consistently across various patient demographics [16].

Healthcare analytics systematically utilizes data and analytical methods to drive decision-making in healthcare settings, leveraging data from EHRs and other sources to enhance risk prediction and patient care [7]. Integrating diverse data types, such as clinical notes and time-series data, provides comprehensive insights into patient health and addresses the complexities of predicting disease risks using multimodal data [10].

2.3 Role of Cardiovascular Risk Factors

Cardiovascular risk factors are pivotal in stroke occurrence and recurrence, necessitating comprehensive understanding to enhance predictive models and patient management strategies. Hypertension, diabetes, and hyperlipidemia are well-established contributors to CVDs, closely associated with increased stroke risk [17]. The interplay among these factors complicates stroke risk, elevating the likelihood of both initial events and recurrences.

Managing these risk factors is complicated by the heterogeneity of medical data across healthcare settings. Federated Learning (FL) offers a promising approach to address data-sharing constraints imposed by privacy regulations, enabling collaborative model training while preserving patient confidentiality [18]. This is particularly relevant for stroke prediction, where diverse datasets can enhance model robustness and accuracy.

Multi-task learning methods in healthcare are similarly constrained by data-sharing limitations, hindering model development that addresses multiple risk factors simultaneously [19]. Overcoming these challenges is vital for improving predictive model precision and tailoring interventions to individual patient profiles. The Tailored Bayes framework exemplifies an innovative approach to managing varying costs associated with classification errors in critical healthcare scenarios, thereby enhancing the reliability of risk predictions [20].

Incorporating a broad range of cardiovascular risk factors into predictive models is essential for accurately assessing stroke recurrence risk. This requires implementing sophisticated computational techniques, such as multi-task and deep learning, alongside a strategic framework for data integration and model development that incorporates domain-specific knowledge and contextual understanding, enhancing interpretability and applicability in real-world healthcare settings [21, 22, 23, 19, 1]. By leveraging innovative frameworks and addressing data-sharing challenges, healthcare professionals can develop more effective strategies for stroke prevention and management, ultimately leading to improved patient outcomes.

3 Machine Learning Techniques in Stroke Risk Prediction

Category	Feature	Method
Supervised and Unsupervised Learning Techniques	Model Combination	XGBoost-EP[24]
Deep Learning Approaches	Hybrid Neural Networks	ICRNN[25]
	Data Fusion Techniques	DLS[6]
Innovative Frameworks and Algorithms	Probabilistic and Censoring Methods	Bayes-AC[26], RFRE.PO[13]
	Dynamic and Adaptive Techniques	DECENT[27], CUSUM[28]
	Advanced Representation Strategies	CCE[16], LERP[7]
	Integration and Multimodal Approaches	DH[15], JFDRP[10]

Table 1: Overview of Machine Learning Methods for Stroke Risk Prediction. This table categorizes various machine learning techniques into supervised and unsupervised learning, deep learning approaches, and innovative frameworks and algorithms. Each category is further detailed with specific methods and their respective references, demonstrating their application in enhancing predictive accuracy and personalized healthcare interventions.

Machine learning techniques have become pivotal in stroke risk prediction, transforming clinical decision-making and improving patient outcomes. Table 1 provides a comprehensive summary of the machine learning methods employed in stroke risk prediction, highlighting the diversity of approaches and their contributions to improving predictive accuracy and patient outcomes. Additionally, Table 2 offers a detailed comparison of the machine learning techniques employed in stroke risk prediction, illustrating the diversity of approaches and their specific applications in enhancing predictive accuracy and personalized healthcare. This section delves into various methodologies, starting with supervised and unsupervised learning, which are foundational for leveraging data-driven insights to predict stroke risk and inform preventative strategies. Figure 2 illustrates the hierarchical categorization

AI-generated, for reference only.

of these machine learning techniques in stroke risk prediction, encompassing not only supervised and unsupervised learning but also deep learning approaches, ensemble and hybrid methods, and innovative frameworks and algorithms. Each category is further detailed with specific techniques and applications, highlighting their roles in enhancing predictive accuracy and contributing to personalized healthcare interventions.

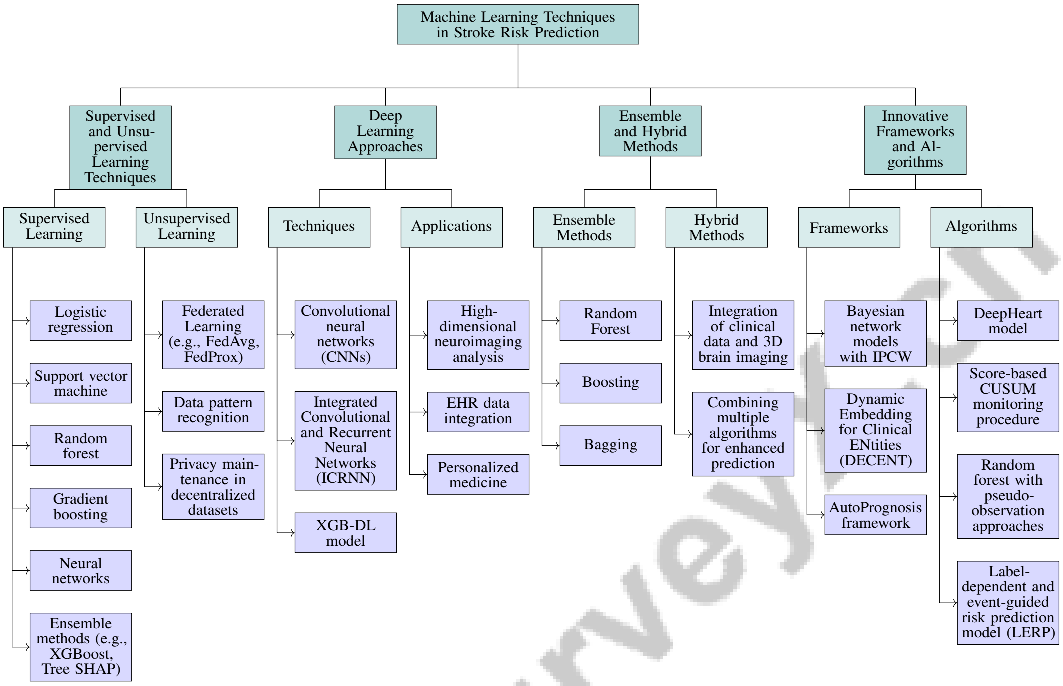

Machine Learning Techniques in Stroke Risk Prediction
Supervised and Unsupervised Learning Techniques
Deep Learning Approaches
Ensemble and Hybrid Methods
Innovative Frameworks and Algorithms
Supervised Learning
Unsupervised Learning
Techniques
Applications
Ensemble Methods
Hybrid Methods
Frameworks
Algorithms
Logistic regression
Support vector machine
Random forest
Gradient boosting
Neural networks
Ensemble methods (e.g., XGBoost, Tree SHAP)
Federated Learning (e.g., FedAvg, FedProx)
Data pattern recognition
Privacy maintenance in decentralized datasets
Convolutional neural networks (CNNs)
Integrated Convolutional and Recurrent Neural Networks (ICRNN)
XGB-DL model
High-dimensional neuroimaging analysis
EHR data integration
Personalized medicine
Random Forest
Boosting
Bagging
Integration of clinical data and 3D brain imaging
Combining multiple algorithms for enhanced prediction
Bayesian network models with IPCW
Dynamic Embedding for Clinical ENtities (DECENT)
AutoPrognosis framework
DeepHeart model
Score-based CUSUM monitoring procedure
Random forest with pseudo-observation approaches
Label-dependent and event-guided risk prediction model (LERP)

Figure 2: This figure illustrates the hierarchical categorization of machine learning techniques in stroke risk prediction, encompassing supervised and unsupervised learning, deep learning approaches, ensemble and hybrid methods, and innovative frameworks and algorithms. Each category is further detailed with specific techniques and applications, highlighting their roles in enhancing predictive accuracy and contributing to personalized healthcare interventions.

3.1 Supervised and Unsupervised Learning Techniques

Supervised learning, which uses labeled datasets, is effective in stroke risk prediction by utilizing techniques like logistic regression, support vector machine, random forest, gradient boosting, and neural networks [5]. These models, optimized with ensemble methods such as XGBoost and Tree SHAP, enhance interpretability and accuracy in cardiovascular risk assessment [24]. Techniques like greedy search algorithms and class balancing further refine predictive performance [29].

Conversely, unsupervised learning reveals hidden patterns in data, crucial in healthcare where labeled data is scarce. Federated Learning (FL) algorithms, such as FedAvg and FedProx, facilitate collaborative model training across decentralized datasets while maintaining privacy [30]. These methods are applicable in both supervised and unsupervised learning for stroke risk prediction, integrating diverse data sources.

The empirical characterization of fairness in machine learning highlights the need for sociotechnical considerations in applying these techniques to stroke risk prediction [31]. The complexity of clinical data necessitates personalized and interpretable machine learning techniques [32]. By addressing data-sharing challenges, healthcare professionals can develop effective strategies for stroke prevention and management, improving patient outcomes.

3.2 Deep Learning Approaches

Deep learning significantly enhances stroke risk prediction by analyzing high-dimensional neuroimaging and EHR data. Techniques like convolutional neural networks (CNNs) integrate multimodal data, including MRI scans and clinical characteristics, improving classification accuracy and enabling personalized medicine [33, 34]. The Integrated Convolutional and Recurrent Neural Networks (ICRNN) model exemplifies this by combining spatial and temporal data analysis for improved predictive accuracy [25].

The XGB-DL model, which integrates feature selection with deep learning, demonstrates the importance of combining feature engineering with deep learning for stroke outcome prediction [6]. Semi-supervised learning methods, as proposed by [15], leverage both labeled and unlabeled data to enhance prediction accuracy, particularly when labeled data is limited.

The integration of convolutional and recurrent neural networks or semi-supervised learning methods in deep learning models effectively predicts stroke risk, providing insights for personalized healthcare interventions [15].

3.3 Ensemble and Hybrid Methods

Ensemble and hybrid methods enhance stroke risk model accuracy by combining multiple algorithms to utilize their strengths. The Random Forest (RF) classifier, for instance, aggregates predictions from decision trees, achieving superior accuracy [29]. Hybrid methods combine machine learning techniques to enhance prediction and inference, crucial in complex domains like biological systems [35, 19].

Ensemble methods like boosting and bagging refine model predictions, optimizing performance in heterogeneous healthcare datasets. These advanced methods enhance predictive accuracy by integrating diverse data modalities, such as clinical data and 3D brain imaging, leading to more reliable stroke risk evaluations and improved patient outcomes [36, 37, 38, 39].

3.4 Innovative Frameworks and Algorithms

Innovative frameworks and algorithms enhance stroke risk prediction by improving accuracy and interpretability. Bayesian network models with Inverse Probability of Censoring Weights (IPCW) exemplify the use of probabilistic approaches for complex healthcare data [26]. The Dynamic Embedding for Clinical ENtities (DECENT) framework learns dynamic embeddings for healthcare entities, improving risk assessments [27].

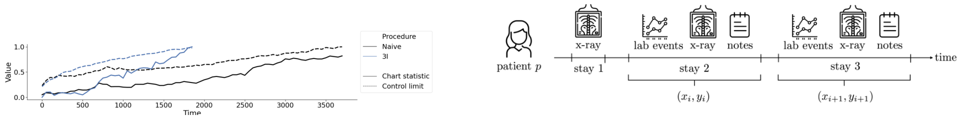
Clinical concept embeddings, as shown by [16], improve predictive performance using advanced representation learning. The DeepHeart model combines wearable technology with deep learning for enhanced stroke risk prediction [15]. The score-based CUSUM monitoring procedure with dynamic control limits addresses challenges from medical interventions, enhancing model robustness [28].

Random forest algorithms with pseudo-observation approaches improve stroke risk prediction by handling censored data and complex interactions [13]. The label-dependent and event-guided risk prediction model (LERP) enhances accuracy and interpretability [7]. The AutoPrognosis framework automates model selection and tuning, discovering new predictors [9]. The Joint-Fusion Framework for Disease Risk Prediction (JFDRP) integrates multiple healthcare data modalities for comprehensive risk assessments [10].

These frameworks and algorithms leverage advanced computational techniques and diverse data sources to enhance stroke risk prediction. By addressing data complexity and representation learning challenges, these approaches, such as self-supervised multimodal frameworks, show potential to improve predictive accuracy and risk assessment in stroke management. Integrating diverse data types, including neuroimaging and EHRs, supports personalized and effective stroke care strategies [40, 38, 41, 36, 34].

As shown in Figure 3, the integration of innovative frameworks and algorithms is crucial in enhancing predictive accuracy and healthcare outcomes in stroke risk prediction. A line graph compares three monitoring methods over time, highlighting the importance of selecting appropriate strategies to track patient health changes over extended periods. A timeline representation of sequential data collection in medical care underscores patient care management's dynamic nature, detailing patient interactions and informing predictive models. These examples illustrate the transformative impact of machine learning techniques and innovative algorithms in refining stroke risk prediction, contributing to more personalized and effective healthcare interventions [42, 10].

AI-generated, for reference only.



(a) Comparison of three methods for monitoring a process over time[42]

(b) Sequential Data Collection in Medical Care: A Timeline of Patient Care[10]

Figure 3: Examples of Innovative Frameworks and Algorithms

Feature	Supervised and Unsupervised Learning Techniques	Deep Learning Approaches	Ensemble and Hybrid Methods
Data Type	Labeled And Unlabeled	Neuroimaging, Ehr	Clinical, 3D Imaging
Model Enhancement	Ensemble Methods	Feature Selection	Boosting, Bagging
Predictive Focus	Stroke Risk	Personalized Medicine	Stroke Risk

Table 2: This table provides a comparative analysis of various machine learning methodologies utilized in stroke risk prediction. It categorizes the techniques into supervised and unsupervised learning, deep learning approaches, and ensemble and hybrid methods, detailing their data types, model enhancement strategies, and predictive focus. The table highlights the diverse applications and contributions of these methods to improving predictive accuracy and personalized medicine in stroke risk assessment.

4 Predictive Modeling and Artificial Intelligence

4.1 Development and Validation of Predictive Models

The development and validation of predictive models for stroke risk involve integrating diverse data sources and employing advanced machine learning (ML) techniques. Deep learning systems (DLS) that incorporate photoplethysmography (PPG) data with demographic information exemplify a comprehensive model development approach, enhancing generalizability across patient populations [6]. Addressing class imbalance is crucial for model reliability and calibration, as shown in benchmarks evaluating ML techniques for cardiovascular disease (CVD) risk prediction [9]. Cross-attention mechanisms, which learn attention weights for medical notes based on semantic similarity to disease risk labels, enhance model interpretability and accuracy [7].

Integrating small language models to fuse various data sources facilitates joint reasoning and enhances predictive accuracy [10]. Effective communication of predictive model outputs to patients is equally important; Linguistic Uncertainty Communication (LUC) conveys uncertainty in AI risk predictions using natural language, fostering patient understanding and trust [43]. Validation of predictive models requires addressing challenges related to reliability and applicability, with a focus on transparency and reliability, particularly in predicting patient deterioration and hospital readmissions [7]. Score-based approaches to monitor model performance over time, while accounting for biases introduced by treatment decisions, support continuous improvement and adaptation.

Advanced ML techniques, such as support vector machines, Random Forest, and deep learning models like convolutional and recurrent neural networks, enhance stroke prevention and management strategies by integrating high-dimensional neuroimaging data with clinical and demographic information. Ensuring model transparency and fairness is essential for validating and optimizing these approaches, ultimately improving post-stroke recovery predictions and patient outcomes [36, 34]. Successful model development and validation depend on integrating diverse data sources, enhancing interpretability, and effectively communicating predictions to patients.

4.2 Integration into Clinical Practice

Integrating predictive models into clinical practice requires addressing technical, operational, and ethical dimensions. A robust governance model is essential for AI model development and evaluation, emphasizing public and patient involvement to ensure alignment with patient needs and ethical standards [44]. Efficient and scalable deployment of predictive models is critical, with the Serverless on FHIR architecture, utilizing containerized microservices and serverless computing, offering a cloud-based solution for deploying ML models in healthcare environments [45]. This architecture ensures models remain accessible and up-to-date, enhancing clinical utility.

AI-generated, for reference only.

Reliability in clinical practice demands rigorous development and evaluation methodologies to ensure accurate risk assessments. Techniques such as shrinkage methods mitigate overfitting, while language models enhance interpretation of structured Electronic Health Records (EHRs) and ensure explainability, fostering trust among healthcare professionals. Integrating domain-relevant evidence into predictions further enhances model reliability and reduces diagnostic errors, highlighting the necessity for comprehensive validation and transparency in the predictive modeling process [40, 46, 47, 48, 37].

The governance model underscores the importance of public and patient involvement in AI model development and evaluation [49]. Engaging patients and the public enhances the transparency and trustworthiness of predictive models, essential for successful clinical implementation [50]. Leveraging diverse data sources and ensuring model reliability are critical for developing comprehensive risk prediction models [7], enabling healthcare professionals to make informed decisions and enhance patient care and outcomes.

4.3 Model Interpretability and Explainability

Model interpretability and explainability are crucial in healthcare analytics, influencing the trust and adoption of machine learning and artificial intelligence (AI) systems by healthcare professionals and patients [4]. Interpretability refers to the ease of understanding decision causes, while explainability provides insights into the model's decision-making process, both essential for ensuring transparency, accountability, and acceptance of AI-driven predictions in clinical practice [43]. The complexity of machine learning models, particularly deep learning architectures, often leads to a "black box" perception, hindering adoption in clinical settings as practitioners may hesitate to trust models they cannot fully comprehend [31].

Recent advancements in explainable AI (XAI) focus on creating domain-appropriate representations that align with clinical workflows and enhance interpretability. Models like XGBoost combined with Tree SHAP provide interpretable outputs, underscoring the importance of model transparency in healthcare [24]. Such approaches enable healthcare professionals to understand the factors driving predictions, facilitating informed decision-making and improving patient outcomes.

The need for model interpretability is further emphasized by the complexities of intersecting group identities in clinical contexts. An intersectional framework for counterfactual fairness highlights the importance of developing models that consider diverse patient needs to ensure equitable health outcomes [4]. This is particularly relevant in stroke risk prediction, where diverse patient populations present varying risk factors and health profiles.

Integrating linguistically appropriate interfaces for communicating risk predictions is essential for model explainability. User-centered design in these interfaces effectively conveys inherent uncertainties in predictive models, enhancing trust and facilitating adoption in clinical practice [43].

4.4 Interpretable and Trustworthy AI Models

The demand for interpretable and trustworthy AI models in medical applications stems from the critical need for transparency and reliability in clinical decision-making. As AI systems become increasingly integrated into healthcare, ensuring that models provide explanations aligned with clinical realities is vital for acceptance and effective use by healthcare professionals [51]. Trustworthy AI models must deliver accurate predictions while offering insights into the underlying factors influencing these predictions, fostering trust and confidence among users.

Benchmark evaluations of popular explainable AI (XAI) methods in clinical settings underscore the necessity for explanations that are both accurate and meaningful to medical practitioners. Achieving congruence between model predictions and clinical realities is essential for the adoption of XAI methods in healthcare, fostering trust and confidence among users [51]. This congruence is particularly critical in stroke risk prediction, where precise and understandable risk assessments are vital for patient management and treatment planning.

Developing interpretable AI models must consider diverse patient needs to ensure equitable health outcomes [4]. Striking a balance between model complexity and interpretability is crucial for clinical acceptance; while complex models may yield high predictive accuracy, they often lack transparency, raising concerns about biases and trustworthiness. Recent advancements in explainable AI emphasize

AI-generated, for reference only.

the importance of interpretability in clinical risk prediction, highlighting that models must provide clear decision-making logic to stakeholders. Tools like GAM Changer facilitate this process by allowing domain experts to interactively edit models, aligning them with human knowledge and values, ultimately enhancing reliability and clinical utility [48, 23]. By focusing on interpretable and trustworthy AI models, healthcare providers can enhance the effectiveness of AI-driven interventions, leading to improved patient outcomes and more personalized care.

5 Healthcare Analytics and Data Utilization

5.1 Types of Data Utilized in Healthcare Analytics

Healthcare analytics employs various data sources to enhance risk prediction models, especially for stroke recurrence. Central to this are Electronic Health Records (EHRs), which offer structured patient data, including demographics, lab results, and prescriptions [7, 52]. Integrating EHRs with imaging, clinical notes, and lab event data provides a holistic view of patient health, significantly boosting predictive capabilities [10].

Multimodal data integration is vital for advancing stroke risk prediction, incorporating clinical notes, medical imaging, and time-series data to develop more accurate models [53]. Advanced computational techniques facilitate insights from complex datasets, leading to better patient outcomes and strategies. AgentMD, a language agent synthesizing multimodal data, underscores the importance of diverse data sources in patient monitoring and care [54, 8]. Hierarchical transformer-based models further enhance EHR data insights, tackling high-dimensional and heterogeneous healthcare data challenges [25, 54]. By leveraging healthcare analytics and diverse data sources, professionals can improve stroke risk prediction and implement effective preventative strategies, emphasizing a data-driven approach in healthcare [53].

5.2 Data Quality and Integration Challenges

Data quality and integration pose significant challenges in healthcare analytics for stroke risk prediction. Imbalances in healthcare datasets often underrepresent patient groups, leading to biased predictions and compromising model accuracy and fairness [55, 56, 4, 31, 2]. Traditional methods for correcting class imbalances are often inadequate, especially for small or intersecting subgroups, necessitating fairness-focused model development.

Integrating multimodal data from EHRs, imaging, and genetic data adds complexity and heterogeneity challenges [10]. Effective data integration is crucial for enhancing model accuracy and comprehensive risk assessments. Achieving fairness while maintaining performance is critical, particularly in stroke risk prediction where data quality disparities can bias outcomes [31]. Federated learning (FL) offers a solution by enabling collaborative model training across decentralized datasets while preserving privacy [18]. This approach allows leveraging diverse data without compromising confidentiality, facilitating robust predictive model development.

Advanced data processing techniques, such as multi-task learning, hold promise for improving model precision and reliability [19]. Developing interpretable models is essential for trust among healthcare professionals and patients [31]. Innovative approaches like federated learning and multi-task learning enable better utilization of diverse patient data for accurate and equitable stroke risk prediction.

Integrating multimodal healthcare data, including clinical notes and real-time interactions, is crucial for comprehensive predictive models capturing disease risk complexities [10]. Advanced techniques, such as dynamic embedding and probabilistic modeling, improve data integration and analysis, enhancing stroke risk model predictive power [27]. Innovative frameworks and algorithms address data quality and integration challenges in stroke risk prediction, leveraging advanced computational techniques and overcoming data-sharing constraints to improve patient outcomes. Integrating diverse data sources like EHRs and medical ontologies with sophisticated predictive models is crucial for optimizing stroke risk prediction and patient care, though challenges like data privacy and validation remain [36, 40].

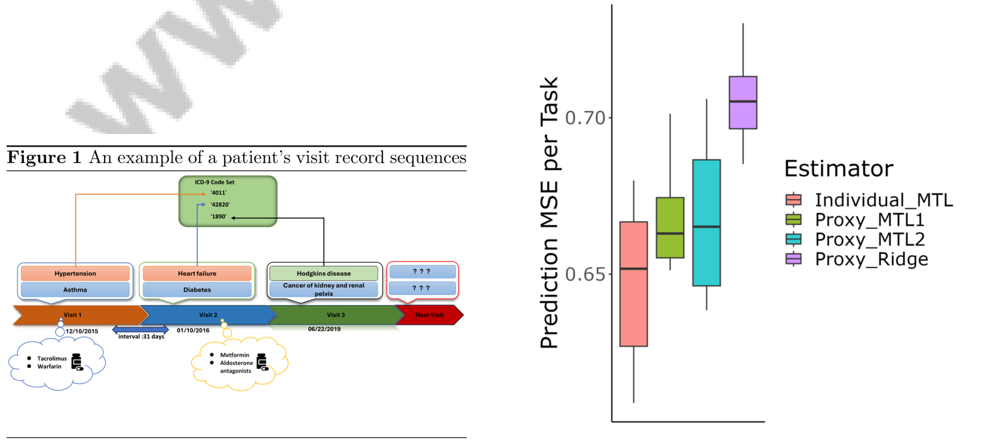
5.3 Advanced Data Processing Techniques

The development of stroke risk prediction models increasingly relies on advanced data processing techniques, including machine learning and multimodal approaches that manage healthcare data's complexity from sources like EHRs and social media. Methods such as support vector machines, recurrent neural networks, and language models enhance predictive accuracy by integrating structured and unstructured data while addressing privacy and computational constraints [46, 40, 39, 36, 57]. The RECAP-KG framework constructs knowledge graphs from primary care notes, enhancing model interpretability and accuracy.

Dynamic embedding techniques, like the DECENT framework, capture patient data's temporal dynamics [27]. Score-based cumulative sum (CUSUM) monitoring procedures with dynamic control limits address confounding medical interventions in stroke risk prediction, enabling timely interventions [28]. The integration of small language models for joint reasoning on multimodal data sources marks a significant advancement in healthcare analytics, enhancing stroke risk assessments' predictive accuracy [10]. Advanced processing techniques, including dynamic embeddings and probabilistic models, underscore innovative frameworks' importance in improving stroke risk prediction by integrating diverse data sources while modeling complex interactions among patients, healthcare professionals, and resources. This comprehensive approach enhances predictive accuracy for various healthcare outcomes and improves AI-driven insights' interpretability, leading to better patient outcomes and more efficient healthcare delivery [58, 46, 27, 22, 32].

Ensemble and hybrid methods, particularly those integrating convolutional neural networks with ensemble techniques, enhance stroke risk model accuracy by leveraging diverse data sources, including 3D brain imaging and clinical data. This approach captures complementary information and addresses traditional methods' limitations, leading to more precise stroke risk assessments. Studies show such models outperform existing unimodal and multimodal methods, demonstrating significant improvements in metrics like ROC-AUC and balanced accuracy, paving the way for effective clinical applications in stroke prevention [36, 38]. By harnessing multiple algorithms' strengths, these methods enhance model robustness and adaptability, crucial for handling heterogeneous and high-dimensional healthcare datasets.

Advanced computational techniques, such as Bayesian network models incorporating Inverse Probability of Censoring Weights (IPCW) and the Tailored Bayes framework, highlight their potential in improving stroke risk prediction. These methods leverage large, heterogeneous datasets from EHRs to address complexities of right-censored data and unequal misclassification costs, ultimately enhancing predictive accuracy compared to traditional models like the Cox proportional hazards model [20, 26, 59]. Employing these methodologies enhances predictive model accuracy and interpretability, improving patient care and outcomes in stroke management.



(a) An example of a patient's visit record sequences[60]

(b) Proxy MTL Estimators for Predicting the MSE per Task[19]

Figure 4: Examples of Advanced Data Processing Techniques

AI-generated, for reference only.

As illustrated in Figure 4, advanced data processing techniques are pivotal for extracting meaningful insights from complex healthcare datasets. The first example depicts a patient's visit record sequences, showcasing the chronological order of medical conditions such as hypertension, asthma, and heart failure, aiding in understanding health progression over time for informed clinical decision-making. The second example highlights the use of Proxy Multi-Task Learning (MTL) estimators to predict the Mean Squared Error (MSE) per task, comparing four estimators—Individual_MTL, Proxy_MTL1, Proxy_MTL2, and Proxy_Ridge. Analyzing the prediction MSE allows healthcare professionals to better assess model performance, enhancing the accuracy of healthcare predictions and interventions. These examples underscore the transformative potential of advanced data processing techniques in optimizing healthcare outcomes and resource allocation [60, 19].

6 Challenges and Limitations

The integration of artificial intelligence (AI) in healthcare, particularly for stroke risk prediction, faces complex ethical and privacy challenges that impact its clinical acceptance and effectiveness. Addressing these challenges is essential for the responsible deployment of AI technologies. This section explores the ethical and privacy considerations necessary to ensure equitable clinical outcomes.

6.1 Ethical and Privacy Concerns

AI in healthcare introduces significant ethical and privacy issues, particularly the risk of reinforcing existing biases within healthcare systems, often rooted in training data reflecting historical disparities in healthcare access and outcomes [31]. Current fairness methods inadequately address intersecting social identities, potentially leading to biased predictions that disadvantage specific patient groups [4]. Robust data governance frameworks are crucial to protect patient information, as the reliance on electronic health records (EHRs) and sensitive data raises privacy concerns [8]. Federated Learning (FL) offers a promising solution by enabling collaborative model training on decentralized datasets while preserving patient confidentiality [30]. Additionally, explainable AI (XAI) methods must resonate with healthcare professionals to foster trust in AI-driven outcomes [51].

6.2 Model Complexity and Interpretability

The complexity of machine learning (ML) models presents challenges to interpretability in stroke risk prediction, with "black box" models often hindering trust and utilization by healthcare professionals [51]. Balancing model complexity with interpretability is crucial, as complex models may achieve high predictive accuracy but lack transparency [15]. Managing non-linear relationships and confounding factors adds complexity [28], and fairness metrics often overlook individual-level discrimination [56]. The gap between AI predictions and patient comprehension complicates interpretability [43]. Developing interpretable AI models that provide clear explanations for predictions is essential for enhancing clinical decision-making and fostering trust among healthcare professionals and patients [40, 51, 48, 22].

6.3 Scalability and Computational Constraints

Scalability and computational constraints are significant challenges for AI in healthcare, particularly for stroke risk prediction. Adequate computing resources are essential, as resource-limited environments can diminish the performance benefits of advanced AI models [41]. The complexity of multi-modal data requires substantial computational power, and operational costs of advanced models like those used by AgentMD further challenge scalability [54]. Federated learning algorithms, such as SCAFFOLD and FedDyn, enhance scalability by enabling collaborative model training across decentralized datasets without sacrificing performance [18]. However, scalability to high-dimensional data remains a challenge, as existing methods often struggle to generalize across diverse patient populations [50]. Optimizing computational resources and developing scalable AI models are necessary to improve stroke risk prediction and healthcare efficiency.

6.4 Generalizability and Bias

Achieving generalizability and addressing bias in predictive models are critical challenges for deploying machine learning (ML) and artificial intelligence (AI) in stroke risk prediction. Reliance on single datasets can limit model generalizability across diverse populations, as demonstrated by the limitations of models evaluated on homogeneous cohorts like the MIMIC-III dataset [7]. Traditional fairness metrics often fail to account for intersecting identities, complicating efforts to ensure equitable predictions [4]. This issue is exacerbated by assumptions of conditional independence among outcomes, leading to inaccurate joint risk estimates [14]. Class imbalance corrections can result in miscalibrated models, underscoring the need for robust validation methodologies to ensure reliable performance across datasets [2]. A holistic strategy that integrates diverse data sources and advanced machine learning techniques, while engaging with clinical experts, is essential for enhancing the usability and applicability of predictive models in medical decision-making [22, 61, 32]. Prioritizing equitable and transparent AI systems will improve the reliability and applicability of predictive models, advancing stroke risk prediction and patient outcomes.

7 Case Studies and Applications

In the context of advancing our understanding of stroke risk prediction, it is essential to explore the practical implications of machine learning (ML) and artificial intelligence (AI) through various case studies and real-world applications. These examples not only illustrate the theoretical frameworks discussed in the preceding sections but also demonstrate how innovative methodologies can be effectively implemented in clinical settings. The following subsection delves into specific case studies that highlight the successful integration of ML and AI in stroke risk prediction, showcasing their potential to improve predictive accuracy and enhance patient outcomes.

7.1 Case Studies and Real-World Applications

The application of machine learning (ML) and artificial intelligence (AI) in stroke risk prediction has been exemplified through various case studies and real-world implementations, demonstrating their potential to enhance predictive accuracy and improve patient outcomes. A notable example is the study by [62], which illustrates the successful use of ML in stroke risk prediction without relying on advanced imaging techniques. This approach highlights the capability of ML models to achieve high accuracy in predictions by leveraging clinical data, thereby offering a cost-effective solution for stroke risk assessment in diverse healthcare settings.

Another significant case study involves the application of the Dynamic Attention-based Co-clustering Method (DACM), which was tested on benchmark datasets against several baseline methods, including Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN), under varying noise conditions [63]. The results demonstrated the robustness of DACM in handling noisy data, underscoring its potential for real-world applications where data quality and consistency are often challenging.

The study by [37] further exemplifies the use of ML in predicting mortality risk in patients with cardiovascular disease. Conducted on a clinical dataset of 3,728 patients from the Massachusetts General Hospital, this research utilized electrocardiogram and tabular data to enhance prediction accuracy. The integration of diverse data sources in this study highlights the importance of comprehensive data utilization in improving the reliability of predictive models.

In the context of interpretability, the Generalized Additive Model (GAM) Changer was evaluated for its application in healthcare models related to pneumonia and sepsis risk predictions, demonstrating its utility in real-world scenarios [23]. This evaluation underscores the critical role of model interpretability in fostering trust and acceptance of AI-driven predictions among healthcare professionals.

Additionally, the application of informative priors to improve the reliability of predictions was assessed using the MIMIC-IV clinical time-series dataset and MIMIC-CXR chest X-ray images for classifying acute care conditions [64]. This approach highlights the potential of integrating prior knowledge into predictive models to enhance their performance and applicability in clinical settings.

AI-generated, for reference only.

The performance of counterfactual prediction models targeting the statin-naïve risk of cardiovascular disease was evaluated using the Multi-Ethnic Study on Atherosclerosis (MESA) [65]. This study underscores the importance of considering diverse patient demographics in developing equitable and reliable predictive models.

Furthermore, innovative approaches such as sentiment analysis of Twitter data have been demonstrated to effectively predict individuals at risk of cardiovascular disease, outperforming traditional demographic data methods [39]. This novel application of social media data highlights the potential of unconventional data sources in enhancing disease risk prediction.

The case studies and real-world applications presented demonstrate the significant transformative potential of machine learning (ML) and artificial intelligence (AI) in stroke risk prediction. They highlight the critical need to integrate diverse data sources—such as neuroimaging, clinical characteristics, and electronic health records—along with innovative methodologies like deep learning and feature selection. These approaches not only enhance predictive accuracy but also aim to improve patient outcomes by addressing challenges such as high data dimensionality and the limitations of traditional predictive models. For instance, advanced convolutional neural networks have shown promising results in classifying post-stroke recovery outcomes, while language models applied to structured electronic health records have improved risk prediction performance. Together, these insights underscore the importance of leveraging cutting-edge technologies to refine risk assessment in clinical practice. [36, 23, 40, 34]

7.2 Comparison of Machine Learning Models

Benchmark	Size	Domain	Task Format	Metric
C-mix[5]	286	Healthcare	Early-readmission Prediction	AUC, C-index
ML-Mortality[29]	100,000	Health Care	Mortality Prediction	AUC
XAI-Bench[51]	1,000	Clinical Predictive Modeling	Risk Prediction	Agreement Metrics
FL-Hetero[18]	5,110	Healthcare	Stroke Prediction	Accuracy, F1-score
CLIP+ITM[38]	5,000	Stroke Risk Prediction	Binary Classification	ROC-AUC, Balanced Accuracy
FTRS[17]	14,293	Neurology	Epidemiological Analysis	Incidence rate, Prevalence rate
ECG-Benchmark[66]	2,300,000	Cardiac Arrhythmia Diagnosis	Multiclass-multilabel Classification	AUPRC, AUROC
ITE[67]	1,000	Health Sciences	Individualized Treatment Effect Prediction	c-for-benefit, calibration slope

Table 3: This table provides a comprehensive overview of representative benchmarks used in stroke risk prediction and related healthcare domains. It details the benchmark name, dataset size, domain of application, task format, and evaluation metrics, offering insights into their scope and applicability in clinical predictive modeling.

In the realm of stroke risk prediction, various machine learning models have been developed and evaluated for their predictive performance and applicability in clinical settings. A comprehensive comparative analysis of various predictive models is essential for discerning the most effective strategies to enhance patient outcomes, as it not only evaluates prediction performance and variable selection across different settings but also incorporates multidisciplinary insights from information technology, medicine, and ethics, ultimately guiding the implementation of decision support systems that can be practically utilized by healthcare professionals. [58, 5, 46, 41, 32] Table 3 presents a detailed comparison of various benchmarks utilized in machine learning models for stroke risk prediction, highlighting their significance in enhancing predictive accuracy and clinical applicability.

Supervised learning algorithms, such as logistic regression, support vector machines (SVM), random forests, gradient boosting, and neural networks, have been widely adopted in stroke risk prediction due to their ability to leverage labeled datasets for training [5]. Among these, random forest classifiers have demonstrated superior predictive accuracy, particularly when augmented with additional features, achieving an Area Under the Curve (AUC) of 0.828, which is significantly higher than traditional methods [29].

Ensemble methods, such as boosting and bagging, have also shown promise in enhancing the performance of predictive models by combining the strengths of multiple algorithms. The application of XGBoost, a popular ensemble method, in conjunction with Tree SHAP for feature importance analysis, underscores the importance of ensemble approaches in improving model interpretability and accuracy [24]. By leveraging the complementary strengths of different models, ensemble methods can achieve superior predictive accuracy and robustness, making them well-suited for stroke risk prediction.

In addition to traditional supervised learning techniques, deep learning approaches have gained traction in stroke risk prediction due to their ability to automatically learn complex patterns from data. Recent advancements in stroke risk prediction have been significantly bolstered by models such as the Integrated Convolutional and Recurrent Neural Networks (ICRNN) and the XGB-DL model. These models leverage sophisticated feature selection techniques alongside deep learning architectures to enhance predictive accuracy. The ICRNN, for instance, effectively captures both short- and long-term temporal patterns in Electronic Health Records (EHRs) while adeptly managing the inherent challenges of missing data without resorting to imputation. Similarly, the XGB-DL model integrates structured clinical data with unstructured medical texts, utilizing a multimodal approach that combines various data types to improve risk assessment outcomes. Collectively, these innovative methodologies represent a promising direction for achieving more precise and personalized stroke risk predictions. [33, 25, 36, 57, 34]. These models underscore the potential of deep learning to improve predictive performance by capturing intricate data patterns and interactions.

The integration of ensemble methods with deep learning architectures, particularly the combination of convolutional neural networks (CNNs) and ensemble techniques, highlights the significant potential of hybrid approaches in enhancing stroke risk prediction. This synergy allows for the accommodation of diverse data modalities, including clinical data and imaging, thereby improving predictive accuracy and enabling more personalized risk assessments. Studies have demonstrated that such hybrid models not only outperform traditional methods but also effectively leverage the strengths of both ensemble learning and deep learning to capture complex patterns in patient data, ultimately leading to better clinical outcomes. [38, 33, 36, 15, 37]. By leveraging the complementary strengths of various models, hybrid methods can achieve superior predictive accuracy and robustness, ultimately leading to more reliable stroke risk assessments and improved patient outcomes.

The comparative analysis of different machine learning models for stroke risk prediction highlights the importance of model interpretability and explainability in healthcare settings. Popular explainable AI (XAI) methods, such as Tree SHAP, have been evaluated for their application in clinical settings, emphasizing the need for transparent and interpretable models that provide meaningful insights to healthcare professionals [51]. Ensuring that predictive models are both accurate and interpretable is crucial for their successful integration into clinical practice.

7.3 Real-World Data Applications

The utilization of real-world data in the development of predictive models for stroke risk prediction is critical for enhancing their applicability and reliability in clinical settings. "Real-world data, which includes a wide array of sources such as electronic health records (EHRs), medical imaging, and clinical notes, offers a holistic perspective on patient health. This comprehensive dataset enhances the development of predictive models by integrating structured and unstructured information, thereby improving accuracy and robustness. For instance, recent studies have demonstrated the efficacy of multimodal approaches that combine structured clinical variables with unstructured medical texts, leading to superior performance in risk prediction tasks. Additionally, leveraging advanced techniques such as language models can facilitate the incorporation of domain knowledge, enabling these models to adapt to various medical vocabularies and handle previously unseen concepts, ultimately resulting in more effective healthcare interventions." [40, 57, 7, 68]

The integration of multimodal healthcare data, including clinical notes, medical imaging, and time-series data, is essential for providing comprehensive insights into patient health and addressing the complexities of predicting disease risks using multimodal data [10]. By leveraging advanced computational techniques, healthcare analytics can extract valuable insights from these diverse data sources, ultimately leading to improved patient outcomes and more effective healthcare strategies.

The use of real-world data in developing predictive models is exemplified by the application of the XGB-DL model, which integrates advanced feature selection techniques with deep learning to enhance predictive accuracy in stroke outcome predictions . This model underscores the promising synergy of traditional feature engineering and advanced deep learning techniques, particularly convolutional neural networks, to enhance stroke risk prediction by effectively integrating high-dimensional neuroimaging data with clinical features, thereby optimizing predictive accuracy in real-world healthcare applications. [36, 33, 66, 34]

The challenges of data quality and integration in healthcare analytics are further addressed through the application of advanced data processing techniques, such as dynamic embedding and probabilistic models . These advanced techniques facilitate the comprehensive integration and analysis of various data sources, including neuroimaging, electronic health records, and social media sentiment, significantly enhancing the predictive accuracy of stroke risk models. By leveraging machine learning algorithms such as convolutional neural networks and large language models, these methods can incorporate a broader range of patient-specific variables, leading to improved identification of at-risk individuals and ultimately resulting in better patient outcomes through personalized care strategies. [36, 39, 40, 34]

The use of real-world data in developing predictive models is particularly relevant in the context of stroke risk prediction, where diverse patient populations present varying risk factors and health profiles. By leveraging advanced computational techniques and integrating multimodal healthcare data, researchers can develop more accurate and equitable predictive models that address the complexities of stroke risk prediction . This approach highlights the critical need for a data-driven methodology in healthcare analytics, as evidenced by the potential to significantly reduce diagnostic errors through enhanced access to patient information in Electronic Health Records (EHRs). By utilizing advanced models like Neural Additive Models and language models (LMs), healthcare providers can generate individualized risk assessments and improve the accuracy of diagnoses. This ultimately leads to better patient outcomes and the development of more effective healthcare strategies, as demonstrated by improved predictive performance in various risk assessment tasks, including mortality risk and adverse event predictions. [46, 7, 27, 40]

7.4 Real-World Data Applications

The utilization of real-world data is pivotal in the development of predictive models for stroke risk prediction, as it provides a comprehensive and nuanced understanding of patient health that can enhance model accuracy and applicability in clinical settings. Real-world data comprises a diverse array of sources, including electronic health records (EHRs), which integrate structured and unstructured data such as clinical notes, laboratory test results, and medical imaging. Each of these sources provides distinct insights into patient conditions and disease progression; for instance, clinical notes contain rich, narrative information that can be systematically analyzed to capture chronological events and enhance risk prediction models. Recent advancements in machine learning have leveraged this multifaceted data, employing techniques like multimodal neural networks that combine textual and structured data to improve outcomes in areas such as cardiovascular risk prediction and disease risk assessment, ultimately leading to more accurate and interpretable healthcare insights. [58, 57, 7]

The integration of multimodal healthcare data is crucial for capturing the multifaceted nature of stroke risk. By combining structured data from EHRs with unstructured data such as clinical notes, researchers can develop predictive models that offer a more holistic view of patient health [10]. This comprehensive approach allows for the identification of complex interactions and patterns that may not be apparent when analyzing data from a single source.

Advanced computational techniques, including dynamic embedding methods like DECENT and probabilistic models for risk prediction, are crucial for effectively processing and analyzing complex real-world data, particularly in healthcare. These techniques enable the extraction of structured representations from heterogeneous data streams, such as electronic health records and patient interactions, facilitating improved predictive modeling for various clinical outcomes. For instance, DECENT has demonstrated significant gains in predicting patient mortality risk and adverse events, while probabilistic models can provide domain-relevant evidence to enhance trust in predictions made by healthcare professionals. [33, 27, 37, 23]. These methods facilitate the effective integration of diverse data sources, enabling the development of robust predictive models that can accurately assess stroke risk across different patient populations . The use of real-world data in predictive modeling is exemplified by the XGB-DL model, which combines feature selection techniques with deep learning to enhance the accuracy of stroke outcome predictions .

The challenges associated with data quality and integration in healthcare analytics are addressed through innovative frameworks that leverage real-world data to improve model performance. For

AI-generated, for reference only.

instance, the application of dynamic embedding techniques allows for the modeling of temporal dynamics in patient data, enhancing the predictive accuracy of stroke risk models [27]. By employing these advanced techniques, researchers can overcome the limitations of traditional data processing methods and develop predictive models that are both accurate and generalizable.

8 Conclusion

This survey underscores the pivotal role of machine learning (ML) and artificial intelligence (AI) in advancing stroke risk prediction, highlighting their potential to enhance predictive precision and patient care. The amalgamation of diverse data sources, such as electronic health records (EHRs) and multimodal healthcare data, is essential for crafting comprehensive predictive models that adeptly capture the intricate interplay of risk factors. Advanced ML methodologies, including supervised, unsupervised, and deep learning techniques, have demonstrated superior performance over traditional approaches, enabling more accurate risk evaluations.

The development of innovative frameworks and algorithms, exemplified by models like the Dynamic Embedding for Clinical ENtities (DECENT) and XGB-DL, has shown significant promise in elevating both the predictive capability and interpretability of stroke risk models. These advancements emphasize the necessity of utilizing state-of-the-art computational strategies and integrating varied data sources to refine stroke risk prediction and enhance patient outcomes.

Model interpretability and explainability are paramount, fostering trust and acceptance among healthcare providers and patients. Ensuring transparency, equity, and impartiality in AI systems is crucial for their successful clinical deployment. The swift integration of automated decision systems in healthcare could potentially amplify existing health disparities, necessitating vigilant oversight and evaluation of these technologies.

Future research should focus on improving the adaptability of predictive models to diverse patient profiles and expanding data sources, such as medication records and laboratory results, to refine predictions. Prioritizing multicenter validations and exploring policy learning are recommended to broaden the framework for assessing individualized treatment effect prediction models. Moreover, incorporating non-linear relationships and interactions within clinical risk prediction models could substantially enhance predictive accuracy and robustness.

Establishing robust validation methodologies is vital to ensure consistent model performance across various datasets and patient demographics. Additionally, future investigations should explore techniques that maintain efficacy in the absence of confounder data and evaluate the applicability of causality-aware approaches across different domains.

References

[1] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, 2021.

[2] Alex Carriero, Kim Luijken, Anne de Hond, Karel GM Moons, Ben van Calster, and Maarten van Smeden. The harms of class imbalance corrections for machine learning based prediction models: a simulation study, 2024.

[3] Simon Meyer Lauritsen, Bo Thiesson, Marianne Johansson Jørgensen, Anders Hammerich Riis, Ulrick Skipper Espelund, Jesper Bo Weile, and Jeppe Lange. The consequences of the framing of machine learning risk prediction models: Evaluation of sepsis in general wards, 2021.

[4] Solvejg Wastvedt, Jared Huling, and Julian Wolfson. An intersectional framework for counterfactual fairness in risk prediction, 2023.

[5] Simon Bussy, Raphaël Veil, Vincent Looten, Anita Burgun, Stéphane Gaïffas, Agathe Guilloux, Brigitte Ranque, and Anne-Sophie Jannot. Comparison of methods for early-readmission prediction in a high-dimensional heterogeneous covariates and time-to-event outcome framework, 2018.

[6] Wei-Hung Weng, Sebastien Baur, Mayank Daswani, Christina Chen, Lauren Harrell, Sujay Kakarmath, Mariam Jabara, Babak Behsaz, Cory Y. McLean, Yossi Matias, Greg S. Corrado, Shravya Shetty, Shruthi Prabhakara, Yun Liu, Goodarz Danaei, and Diego Ardila. Predicting cardiovascular disease risk using photoplethysmography and deep learning, 2023.

[7] Shuai Niu, Yunya Song, Qing Yin, Yike Guo, and Xian Yang. Label-dependent and event-guided interpretable disease risk prediction using ehrs, 2022.

[8] Mitchell Burger. The risk to population health equity posed by automated decision systems: A narrative review, 2022.

[9] Ahmed M Alaa, Thomas Bolton, Emanuele Di Angelantonio, James HF Rudd, and Mihaela Van der Schaar. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 uk biobank participants. *PloS one*, 14(5):e0213653, 2019.

[10] Liv Björkdahl, Oskar Pauli, Johan Östman, Chiara Ceccobello, Sara Lundell, and Magnus Kjellberg. Towards holistic disease risk prediction using small language models, 2024.

[11] Laleh Jalali, Hsiu-Khuern Tang, Richard H. Goldstein, and Joaqun Alvarez Rodrguez. Predicting clinical deterioration in hospitals, 2021.

[12] Juan C. Quiroz, David Brieger, Louisa Jorm, Raymond W Sy, Benjumin Hsu, and Blanca Gallego. Predicting adverse outcomes following catheter ablation treatment for atrial fibrillation, 2023.

[13] Abigail Loe, Susan Murray, and Zhenke Wu. Random forest for dynamic risk prediction or recurrent events: A pseudo-observation approach, 2025.

[14] Glen P. Martin, Matthew Sperrin, Kym I. E. Snell, Iain Buchan, and Richard D. Riley. Clinical prediction models to predict the risk of multiple binary outcomes: a comparison of approaches, 2020.

[15] Brandon Ballinger, Johnson Hsieh, Avesh Singh, Nimit Sohoni, Jack Wang, Geoffrey H. Tison, Gregory M. Marcus, Jose M. Sanchez, Carol Maguire, Jeffrey E. Olgin, and Mark J. Pletcher. Deepheart: Semi-supervised sequence learning for cardiovascular risk prediction, 2018.

[16] Spiros Denaxas, Pontus Stenetorp, Sebastian Riedel, Maria Pikoula, Richard Dobson, and Harry Hemingway. Application of clinical concept embeddings for heart failure prediction in uk ehr data, 2018.

[17] Jona T Stahmeyer, Sarah Stubenrauch, Siegfried Geyer, Karin Weissenborn, and Sveja Eberhard. The frequency and timing of recurrent stroke: an analysis of routine health insurance data. *Deutsches Ärzteblatt International*, 116(42):711, 2019.

[18] Usevalad Milasheuski, Luca Barbieri, Bernardo Camajori Tedeschini, Monica Nicoli, and Stefano Savazzi. On the impact of data heterogeneity in federated learning environments with application to healthcare networks, 2024.

[19] Parker Knight and Rui Duan. Multi-task learning with summary statistics, 2024.

[20] Solon Karapanagiotis, Umberto Benedetto, Sach Mukherjee, Paul D. W. Kirk, and Paul J. Newcombe. Tailored bayes: a risk modelling framework under unequal misclassification costs, 2021.

[21] Jing Mei, Eryu Xia, Xiang Li, and Guotong Xie. Developing knowledge-enhanced chronic disease risk prediction models from regional ehr repositories, 2017.

[22] Shruthi Chari, Prithwish Chakraborty, Mohamed Ghalwash, Oshani Seneviratne, Elif K. Eyigoz, Daniel M. Gruen, Fernando Suarez Saiz, Ching-Hua Chen, Pablo Meyer Rojas, and Deborah L. McGuinness. Leveraging clinical context for user-centered explainability: A diabetes use case, 2021.

[23] Zijie J. Wang, Alex Kale, Harsha Nori, Peter Stella, Mark E. Nunnally, Duen Horng Chau, Mihaela Vorvoreanu, Jennifer Wortman Vaughan, and Rich Caruana. Interpretability, then what? editing machine learning models to reflect human knowledge and values, 2022.

[24] Maria Athanasiou, Konstantina Sfrintzeri, Konstantia Zarkogianni, Anastasia C. Thanopoulou, and Konstantina S. Nikita. An explainable xgboost-based approach towards assessing the risk of cardiovascular disease in patients with type 2 diabetes mellitus, 2020.

[25] Yuxi Liu, Shaowen Qin, Antonio Jimeno Yepes, Wei Shao, Zhenhao Zhang, and Flora D. Salim. Integrated convolutional and recurrent neural networks for health risk prediction using patient journey data with many missing values, 2022.

[26] Sunayan Bandyopadhyay, Julian Wolfson, David M. Vock, Gabriela Vazquez-Benitez, Gediminas Adomavicius, Mohamed Elidrisi, Paul E. Johnson, and Patrick J. O'Connor. Data mining for censored time-to-event data: A bayesian network model for predicting cardiovascular risk from electronic health record data, 2014.

[27] Hankyu Jang, Sulyun Lee, D. M. Hasibul Hasan, Philip M. Polgreen, Sriram V. Pemmaraju, and Bijaya Adhikari. Dynamic healthcare embeddings for improving patient care, 2023.

[28] Jean Feng, Alexej Gossmann, Gene Pennello, Nicholas Petrick, Berkman Sahiner, and Romain Pirracchio. Monitoring machine learning (ml)-based risk prediction algorithms in the presence of confounding medical interventions, 2023.

[29] Maggie Makar, Marzyeh Ghassemi, David Cutler, and Ziad Obermeyer. Short-term mortality prediction for elderly patients using medicare claims data, 2017.

[30] Nathan C. Hurley, Erica S. Spatz, Harlan M. Krumholz, Roozbeh Jafari, and Bobak J. Mortazavi. A survey of challenges and opportunities in sensing and analytics for cardiovascular disorders, 2019.

[31] Stephen R. Pfohl, Agata Foryciarz, and Nigam H. Shah. An empirical characterization of fair machine learning for clinical risk prediction, 2021.

[32] Roland Roller, Klemens Budde, Aljoscha Burchardt, Peter Dabrock, Sebastian Möller, Bilgin Osmanodja, Simon Ronicke, David Samhammer, and Sven Schmeier. When performance is not enough – a multidisciplinary view on clinical decision support, 2022.

[33] Zhengping Che, Yu Cheng, Zhaonan Sun, and Yan Liu. Exploiting convolutional neural network for risk prediction with medical feature embedding, 2017.

[34] Adam White, Margarita Saranti, Artur d'Avila Garcez, Thomas M. H. Hope, Cathy J. Price, and Howard Bowman. Predicting recovery following stroke: deep learning, multimodal data and feature selection using explainable ai, 2023.

[35] H Ij. Statistics versus machine learning. *Nat Methods*, 15(4):233, 2018.

AI-generated, for reference only.

[36] Farnoush Shishehbori and Zainab Awan. Enhancing cardiovascular disease risk prediction with machine learning models, 2024.

[37] Aniruddh Raghu, John Guttag, Katherine Young, Eugene Pomerantsev, Adrian V. Dalca, and Collin M. Stultz. Learning to predict with supporting evidence: Applications to clinical risk prediction, 2021.

[38] Camille Delgrange, Olga Demler, Samia Mora, Bjoern Menze, Ezequiel de la Rosa, and Neda Davoudi. A self-supervised model for multi-modal stroke risk prediction, 2024.

[39] Al Zadid Sultan Bin Habib, Md Asif Bin Syed, Md Tanvirul Islam, and Donald A. Adjeroh. Cardiovascular disease risk prediction via social media, 2023.

[40] Angeela Acharya, Sulabh Shrestha, Anyi Chen, Joseph Conte, Sanja Avramovic, Siddhartha Sikdar, Antonios Anastasopoulos, and Sanmay Das. Clinical risk prediction using language models: Benefits and considerations, 2023.

[41] Ricky Sahu, Eric Marriott, Ethan Siegel, David Wagner, Flore Uzan, Troy Yang, and Asim Javed. Introducing the large medical model: State of the art healthcare cost and risk prediction with transformers trained on patient event sequences, 2024.

[42] Jean Feng, Adarsh Subbaswamy, Alexej Gossmann, Harvineet Singh, Berkman Sahiner, Mi-Ok Kim, Gene Pennello, Nicholas Petrick, Romain Pirracchio, and Fan Xia. Designing monitoring strategies for deployed machine learning algorithms: navigating performativity through a causal lens, 2024.

[43] Adarsa Sivaprasad and Ehud Reiter. Linguistically communicating uncertainty in patient-facing risk prediction models, 2024.

[44] Sandeep Reddy, Sonia Allan, Simon Coghlan, and Paul Cooper. A governance model for the application of ai in health care. *Journal of the American Medical Informatics Association*, 27(3):491–497, 2020.

[45] Bell Raj Eapen, Kamran Sartipi, and Norm Archer. Serverless on fhir: Deploying machine learning models for healthcare on the cloud, 2020.

[46] Denis Jered McInerney, William Dickinson, Lucy C. Flynn, Andrea C. Young, Geoffrey S. Young, Jan-Willem van de Meent, and Byron C. Wallace. Towards reducing diagnostic errors with interpretable risk prediction, 2024.

[47] Ben Van Calster, Maarten van Smeden, and Ewout W. Steyerberg. On the variability of regression shrinkage methods for clinical prediction models: simulation study on predictive performance, 2019.

[48] Munib Mesinovic, Peter Watkinson, and Tingting Zhu. Explainable ai for clinical risk prediction: a survey of concepts, methods, and modalities, 2023.

[49] Xi Sheryl Zhang, Fengyi Tang, Hiroko Dodge, Jiayu Zhou, and Fei Wang. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records, 2019.

[50] Peter Schulam and Suchi Saria. Reliable decision support using counterfactual models, 2018.

[51] Aida Brankovic, David Cook, Jessica Rahman, Wenjie Huang, and Sankalp Khanna. Evaluation of popular xai applied to clinical prediction models: Can they be trusted?, 2023.

[52] Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-behrt: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records, 2021.

[53] Rakhilya Lee Mekhtieva, Brandon Forbes, Dalal Alrajeh, Brendan Delaney, and Alessandra Russo. Recap-kg: Mining knowledge graphs from raw gp notes for remote covid-19 assessment in primary care, 2023.

AI-generated, for reference only.

[54] Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, W John Wilbur, Zhe He, Andrew Taylor, Qingyu Chen, and Zhiyong Lu. Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning, 2024.

[55] Solvejg Wastvedt, Jared D Huling, and Julian Wolfson. Counterfactual fairness for small subgroups, 2024.

[56] Stephen Pfohl, Tony Duan, Daisy Yi Ding, and Nigam H. Shah. Counterfactual reasoning for fair clinical risk prediction, 2019.

[57] Ayoub Bagheri, T. Katrien J. Groenhof, Wouter B. Veldhuis, Pim A. de Jong, Folkert W. Asselbergs, and Daniel L. Oberski. Multimodal learning for cardiovascular risk prediction using ehr data, 2020.

[58] Andrey Kormilitzin, Nemanja Vaci, Qiang Liu, Hao Ni, Goran Nenadic, and Alejo Nevado-Holgado. An efficient representation of chronological events in medical texts, 2020.

[59] Julian Wolfson, Sunayan Bandyopadhyay, Mohamed Elidrisi, Gabriela Vazquez-Benitez, Donald Musgrove, Gediminas Adomavicius, Paul Johnson, and Patrick O'Connor. A naive bayes machine learning approach to risk prediction using censored, time-to-event data, 2014.

[60] Shibo Li, Hengliang Cheng, and Weihua Li. Time-aware heterogeneous graph transformer with adaptive attention merging for health event prediction, 2024.

[61] Shruthi Chari, Prasant Acharya, Daniel M. Gruen, Olivia Zhang, Elif K. Eyigoz, Mohamed Ghalwash, Oshani Seneviratne, Fernando Suarez Saiz, Pablo Meyer, Prithwish Chakraborty, and Deborah L. McGuinness. Informing clinical assessment by contextualizing post-hoc explanations of risk prediction models in type-2 diabetes, 2023.

[62] Luis García-Terriza, José L. Risco-Martín, Gemma Reig Roselló, and José L. Ayala. Predictive and diagnosis models of stroke from hemodynamic signal monitoring, 2023.

[63] Ziyang Zhang, Hejie Cui, Ran Xu, Yuzhang Xie, Joyce C. Ho, and Carl Yang. Tacco: Task-guided co-clustering of clinical concepts and patient visits for disease subtyping based on ehr data, 2024.

[64] L. Julian Lechuga Lopez, Tim G. J. Rudner, and Farah E. Shamout. Informative priors improve the reliability of multimodal clinical data classification, 2023.

[65] Christopher B. Boyer, Issa J. Dahabreh, and Jon A. Steingrimsson. Assessing model performance for counterfactual predictions, 2023.

[66] Eran Zvuloni, Jesse Read, Antônio H. Ribeiro, Antonio Luiz P. Ribeiro, and Joachim A. Behar. On merging feature engineering and deep learning for diagnosis, risk-prediction and age estimation based on the 12-lead ecg, 2022.

[67] J Hoogland, O Efthimiou, TL Nguyen, and TPA Debray. Evaluating individualized treatment effect predictions: a model-based perspective on discrimination and calibration assessment, 2023.

[68] Chun-Chieh Liao, Wei-Ting Kuo, I-Hsuan Hu, Yen-Chen Shih, Jun-En Ding, Feng Liu, and Fang-Ming Hung. Ehr-based mobile and web platform for chronic disease risk prediction using large language multimodal models, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.