Stroke Prediction using Bayesian Modeling

Group 10

Zhu Yufan
Metric Number: A0238993J
Email: e0773591@u.nus.edu

Lu Yiting
Metric Number: A0239591U
Email: e0774189@u.nus.edu

Yu Zizhen
Metric Number: A0239997Y
Email: e0774595@u.nus.edu

Abstract

Stroke is a leading cause of disability and mortality worldwide. Our project aims to leverage the capabilities of Bayesian modeling to improve the accuracy and reliability of stroke prediction. Utilizing a comprehensive dataset encompassing a range of health metrics and lifestyle factors, we developed a Bayesian network to identify key predictors of stroke risk. This approach not only offers precise predictive power but also provides a probabilistic understanding of various risk factors, aiding in early detection and prevention strategies. Our results demonstrate the model's efficacy in stroke prediction, showcasing its potential as a valuable tool in healthcare settings. The project also underscores the importance of integrating advanced data analytics and AI in medical diagnosis, paving the way for more informed and effective healthcare interventions.

November 23, 2023

1 Introduction

Globally recognized as a significant health challenge [24], stroke's sudden onset and potential for severe consequences demand early and effective prediction mechanisms [7]. Timely interventions can greatly reduce the severity of strokes, emphasizing the need for robust predictive models [15]. Recent advancements in machine learning and data-driven methodologies open avenues for enhanced prediction capabilities [3]. This project seeks to tap into Bayesian modeling, celebrated for its comprehensive approach of amalgamating prior knowledge and observed data [25]. In doing so, we consider a range of health determinants, from age and gender to hypertension and heart disease [4, 13]. While the ideal in healthcare would be a model embracing every possible risk factor [12], we've narrowed our focus to the dataset at hand, balancing depth with feasibility [17].

2 Related Works

The prediction of stroke, a critical health issue with widespread impact, has garnered considerable attention in medical research. Various statistical and machine-learning models have been deployed to tackle this challenge. A notable example is the Framingham Stroke Risk Profile (FSRP), a seminal algorithm that integrates multiple risk factors to forecast the 10-year risk of stroke [18]. Despite its acclaim and widespread adoption in clinical settings, the FSRP has faced criticism for its limited precision and inability to adequately address the complex interplay between various stroke risk factors [29].

In the realm of machine learning, advanced models like decision trees, random forests, and neural networks have shown considerable promise in enhancing stroke prediction capabilities [1]. These models excel in processing high-dimensional datasets and uncovering intricate relationships between diverse variables, offering a more nuanced understanding of stroke risk. However, they are often critiqued for their "black-box" nature, which obscures the decision-making process, a limitation that can be particularly concerning in the context of healthcare, where interpretability and transparency are crucial [2].

The broader landscape of machine learning in medicine has been thoroughly explored by Mainali et al. [19]. They highlight that commercially available machine learning algorithms have not only been integrated into clinical practices but have also transformed the speed and efficiency of diagnosis. The rapid evolution of deep learning techniques has further revolutionized the clinical application of these tools, leading to the emergence of innovative algorithms that enhance the accuracy of stroke diagnosis and prognostication.

Wang et al. conducted a systematic review to assess the use of machine learning models in predicting stroke outcomes [30]. Their review underscores a growing trend in leveraging machine learning for such predictions. However, they noted a gap in adherence to basic reporting standards for clinical prediction tools and the availability of models for public use and evaluation, indicating a need for more transparent and accessible machine learning solutions in stroke care.

Amidst this landscape, Bayesian modeling stands out for its unique capacity to integrate prior knowledge and deliver probabilistic predictions, a feature highly valuable in healthcare applications [16]. Despite its potential, the application of Bayesian modeling in stroke prediction is not yet widespread. This project endeavors to fill this void by constructing a Bayesian network specifically designed for stroke prediction. Our work aims to contribute meaningfully to the ongoing efforts in this field, addressing some of the challenges highlighted in the existing literature and paving the way for more effective and interpretable stroke prediction models.

3 Methodology

3.1 Data Collection

The dataset utilized in this study is sourced from Kaggle, titled "Stroke Prediction Dataset" [8]. This dataset aims to assist in predicting the likelihood of a patient experiencing a stroke based on various clinical and lifestyle-related attributes.

The dataset comprises 3245 observations, each providing pertinent information about individual patients. The attributes include:

Table 1: Attributes in the Stroke Prediction Dataset

Attribute	Description
id	A unique identifier for each patient.
gender	Categorical variable with values "Male", "Female".
age	Numerical variable indicating the age of the patient.
hypertension	Binary variable, where 0 indicates the absence and 1 indicates the presence of hypertension.
heart_disease	Binary variable, where 0 denotes no heart diseases and 1 signifies the presence of heart disease.
ever_married	Categorical variable with values "No" or "Yes".
work_type	Categorical variable indicating the type of employment - "Govt_job", "Private", or "Self-employed".
Residence_type	Categorical variable with values "Rural" or "Urban".
avg_glucose_level	Numerical variable representing the average glucose level in the patient's blood.
BMI	Numerical variable indicating the patient's body mass index.
smoking_status	Categorical variable with values "formerly smoked", "never smoked", and "smokes".
stroke	Binary target variable, where 1 indicates the patient had a stroke, and 0 indicates they did not.

3.2 Data Cleaning

The initial dataset, obtained from Kaggle, necessitated a meticulous data-cleaning process due to the presence of missing or ambiguous values in several features.

Firstly, to maintain the integrity of the dataset, rows with missing values in the 'BMI' column were eliminated. This step was crucial as 'BMI' is often a significant predictor in health-related studies and its absence could lead to inaccurate model predictions.

Secondly, the dataset included entries with an 'Unknown' smoking status. These entries could potentially introduce noise into the model, thereby affecting its performance. To prevent this, such rows were discarded.

Lastly, inconsistencies in the 'gender' column were rectified by assigning integer values to the gender categories ('Male' as 0 and 'Female' as 1). Furthermore, rows with undefined gender were removed to ensure clarity and consistency in the dataset.

3.3 Feature Transformation

The data underwent several transformations and encodings to prepare it for modeling:

- The 'id' column was dropped to eliminate non-predictive data, as it uniquely identifies patients without contributing to the model's performance.
- 'gender' was cleaned and encoded as binary (Male to 0, Female to 1) to simplify model input and enhance processing. Rows with undefined gender were removed to maintain data integrity.
- 'age' was binned into decades, providing a simplified and generalized feature for easier model interpretation and less sensitivity to small variations.
- Binary columns 'hypertension', 'heart_disease', and 'ever_married' were maintained or encoded ('Yes' as 1 and 'No' as 0 for 'ever_married') to align with the model's binary input requirements.
- 'work_type' was split into 'is_private' and 'is_self_employed' to transform the multi-class categorical data into a binary format, making it more suitable for the model.
- 'Residence_type' was binary encoded (Urban as 1, Rural as 0) for consistency with the model's binary format.
- 'avg_glucose_level' was categorized into low (below 90), normal (90 to 130), and high (above 130) to create a simplified, ordinal feature for better model processing.
- 'BMI' was categorized based on standard classifications: underweight for a BMI less than 18.5, normal weight for a BMI between 18.5 and 24.9, overweight for a BMI between 24.9 and 29.9, and obese

for a BMI of 29.9 and above. This transformation turns a continuous variable into an ordinal one, facilitating model interpretation.

- 'smoking_status' was mapped to ordinal values to convert categorical data into a format more suitable for modeling.
- The target variable 'stroke' was balanced by oversampling rows where 'stroke' is 1 by a factor of seventeen. This addresses the issue of class imbalance, which can otherwise lead to biased model predictions.
- The dataset was shuffled with a fixed seed to ensure randomness and reduce overfitting risk, contributing to the model's generalizability.

3.4 Modeling

3.4.1 Bayesian Modeling

In our pursuit to predict strokes, we employed Bayesian modeling, a probabilistic approach that integrates both prior knowledge and observed data. The structure of our Bayesian network was meticulously designed based on the intricate relationships and dependencies among various health metrics and lifestyle features present in the dataset.

The Bayesian network's architecture integrates both empirical data and established medical knowledge, ensuring its credibility:

The risk of hypertension, heart_disease, and stroke escalates with age, a correlation consistently documented in medical studies [14, 22]. Gender plays a pivotal role in shaping smoking habits and directly correlates with stroke risk, as evidenced by numerous research findings [26, 27]. Established predictors of stroke, such as hypertension, heart_disease, marital status (ever_married), residence_type, employment status (is_private and is_self_employed), avg_glucose_level, and BMI have been duly integrated into our model [9, 21, 28]. Furthermore, a substantiated link between BMI and conditions like hypertension and heart_disease has been considered [11, 10]. Additionally, the model acknowledges the adverse ramifications of smoking_status on hypertension, heart_disease, and stroke risk, drawing from extensive scientific literature [6, 23, 20, 5].

By incorporating these evidence-based relationships into our Bayesian network, we have ensured that our model is not only data-driven but also grounded in scientific knowledge. This lends credibility to our model's structure and reinforces the robustness of the relationships we've introduced. As such, our model serves as a reliable tool for stroke prediction.

The graphical structure of the Bayesian network is illustrated below:

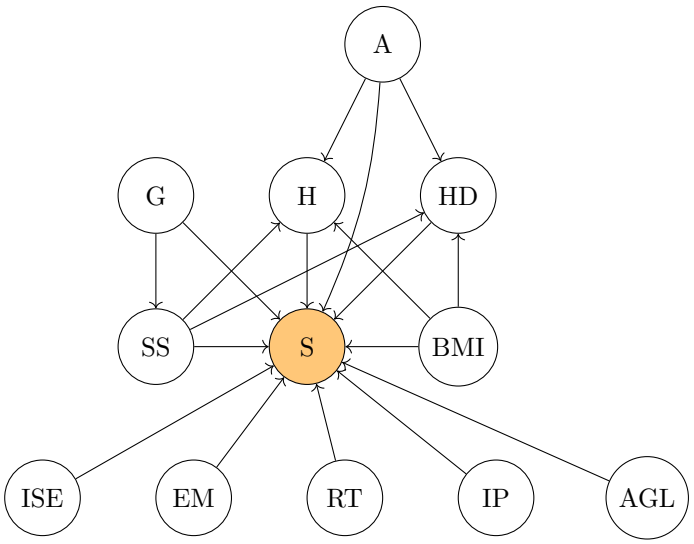

Figure 1: Hierarchical structure of the Bayesian network for stroke prediction.

Legend:

Symbol	Description	Symbol	Description	Symbol	Description
A	age	G	gender	H	hypertension
HD	heart_disease	S	stroke	EM	ever_married
RT	residence_type	AGL	avg_glucose_level	BMI	BMI
SS	smoking_status	IP	is_private	ISE	is_self_employed

Table 2: Legend for the Symbols

Once the structure was defined, we trained the Bayesian model using the Bayesian Estimator with the BDeu prior. After training, the learned Conditional Probability Distributions (CPDs) for each node, such as the "stroke" node, enabled our model to predict stroke likelihood based on various input features.

In addition to our primary Bayesian model, we developed a simpler and more complex version for comparison. The visual representation of the two alternative models is presented in appendix 7 and 5 respectively. The simple model reduces interconnections, while the more complicated one introduces added complexities. The intention behind these variants is to measure their performance against our main model. A thorough analysis of these models is presented in Section 4.2.

3.4.2 Neural Network

In our investigation into predictive analytics for stroke, alongside our primary Bayesian model, we also implemented a deep learning model using TensorFlow and Keras frameworks. This secondary model stands as a point of comparison to ascertain the efficacy of traditional machine learning techniques against the Bayesian approach within the medical prediction field.

The architecture of the model is sequentially layered, beginning with an input layer designed to accept the dimensionality of our feature set. This is followed by two hidden layers, the first comprising 128 neurons and the second 64 neurons, both utilizing the ReLU activation function. The culmination of the model is an output layer with a single neuron employing a sigmoid activation function to output a probability. The model employs the Adam optimizer for the training process and uses binary cross-entropy as the loss function, with accuracy serving as the performance metric. We standardized the data before training and proceeded to fit the model over 10 epochs with a batch size of 32. A visual representation of the neural network is presented in Appendix 3. The efficacy of this model is then assessed on a test dataset to ascertain its generalization capabilities. It offers a benchmark to draw comparisons with our main Bayesian model, illuminating the differences in predictive performance and interpretability inherent to each approach.

4 Performance Evaluation

4.1 Evaluation Metrics

To rigorously gauge the performance of our stroke prediction model, we leverage a suite of evaluation metrics, each offering unique insights into different aspects of model performance:

- **Accuracy**: This is the proportion of total predictions that the model got right. It is the simplest metric for evaluating classification models. Accuracy is particularly intuitive and useful for balanced datasets. However, its utility diminishes with imbalanced datasets, where one class significantly outnumbers the other, as it may give a skewed perception of the model's effectiveness.
- **Precision & Recall**: These metrics provide deeper insights, particularly important in medical applications where the cost of errors can be significant.
  - **Precision**: This metric assesses the proportion of correct identifications. Precision is crucial when the consequences of false positives are severe. For instance, in medical diagnostics, high precision reduces the risk of patients being incorrectly diagnosed with a condition they do not have.
  - **Recall** (also known as Sensitivity): This measures the proportion of actual positives that the model correctly identifies. The high recall is essential in situations where missing a positive case could have dire consequences. For example, in stroke prediction, a high recall rate means the

model is effective at identifying most individuals who are at risk of a stroke, thereby enabling timely intervention.

- **F1-score**: This metric is the harmonic mean of precision and recall. It provides a single score that balances both the concerns of precision and recall, making it particularly useful for imbalanced datasets. The F1-score is effective in scenarios where an equilibrium between false positives and false negatives is essential.
- **ROC Curve and AUC**:
  - **ROC Curve**: A plot showing the performance of a binary classifier at various thresholds, by plotting True Positive Rate (TPR) against False Positive Rate (FPR). It's useful for visualizing and comparing model performance.
  - **AUC** (Area Under the Curve): A single metric ranging from 0 to 1 that summarizes a model's performance across thresholds. Higher AUC indicates better performance, with 1 being perfect accuracy and 0.5 equating to random guessing.

4.2 Comparison of Bayesian Networks with Different Complexities

Detail the differences in performance, interpretability, and other relevant aspects when comparing Bayesian networks of varying complexities.

4.2.1 Model Performance: Proposed Bayesian Network

- **Accuracy**: The proposed model exhibits an impressive accuracy of 97.07%, implying that it correctly predicts the stroke outcome for 97.07% of the test cases.
- **Precision & Recall**:

Metric	No Stroke	Stroke
Precision	1.00 (100%)	0.94 (94%)
Recall	0.94 (94%)	1.00 (100%)

Table 3: Precision and Recall for the Proposed Bayesian Network Model

- **F1-score**: The model has an F1-score of 0.97 and 0.97 for both "No Stroke" and "Stroke" predictions. This harmonized score showcases balanced precision and recall, especially vital given the critical nature of stroke predictions.
- **ROC Curve**:

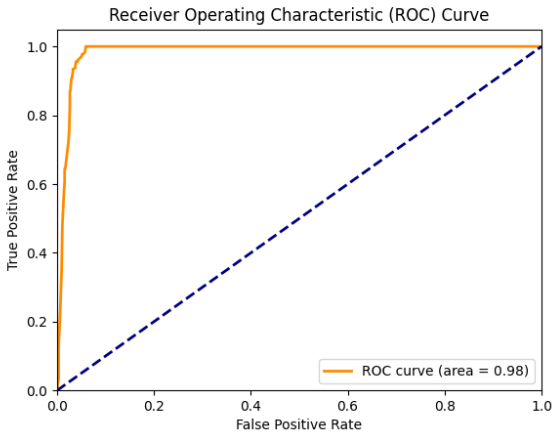

Figure 2: ROC Curve for the Proposed Model

The Receiver Operating Characteristic (ROC) curve provides insights into the trade-off between a true positive rate and a false positive rate. A model with perfect prediction capability will have an Area Under the Curve (AUC) of 1. Our model's AUC is 0.98 suggesting it has an outstanding discriminative power between positive and negative classes.

4.2.2 Model Performance: Simpler Bayesian Network

- **Accuracy**: The simpler model has an accuracy of 90.36%, which is slightly lower than the 97.07% accuracy of the proposed model.
- **Precision & Recall**:

Metric	No Stroke	Stroke
Precision	0.99 (99%)	0.84 (84%)
Recall	0.81 (81%)	0.99 (99%)

Table 4: Precision and Recall for the Simpler Bayesian Network Model

Compared to the proposed model, the simpler model has a slightly lower precision for predicting "Stroke" and a lower recall for predicting "No Stroke". This suggests that while both models are effective, the proposed model is more balanced in terms of precision and recall.

- **F1-score**: The F1-score for the simpler model is slightly lower than that of the proposed model for both "No Stroke" and "Stroke" predictions. This indicates that the proposed model provides a better balance between precision and recall.
- **ROC Curve**:

  The Area Under the Curve (AUC) for the simpler model is 0.95, which is slightly lower than the AUC of 0.98 for the proposed model. This suggests that while both models have excellent discriminative power between positive and negative classes, the proposed model performs slightly better. For a visual representation, see Figure 6 in the Appendix.

4.2.3 Model Performance: More Complex Bayesian Network

- **Accuracy**: The more complex model has an accuracy of 79.80%, which is lower than both the simpler model (90.36%) and the proposed model (97.07%). This suggests that the additional complexity may be leading to overfitting, reducing its ability to generalize to unseen data.
- **Precision & Recall**:

Metric	No Stroke	Stroke
Precision	0.84 (84%)	0.77 (77%)
Recall	0.74 (74%)	0.86 (86%)

Table 5: Precision and Recall for the More Complex Bayesian Network Model

Compared to the proposed model, the more complex model has lower precision and recall for both "No Stroke" and "Stroke" predictions. This indicates that the model may be overfitting to the training data, reducing its predictive performance on the test data.

- **F1-score**: The F1-score for the more complex model is lower than that of both the simpler and proposed models for both "No Stroke" and "Stroke" predictions. This further suggests that the additional complexity may be detrimental to the model's performance.
- **ROC Curve**:

  The Area Under the Curve (AUC) for the more complex model is 0.88, which is lower than both the simpler model (0.95) and the proposed model (0.98). This indicates that while the more complex model still has good discriminative power between positive and negative classes, it is outperformed by the simpler and proposed models. For a visual representation, see Figure 8 in the Appendix.

4.3 Bayesian Network vs. Other Methods

The performance of the Bayesian network is contrasted against other computational methods such as neural networks. Each methodology carries its unique advantages and potential drawbacks, which are significant when considering the application to medical diagnostics like stroke prediction.

4.3.1 Neural Network Performance

The neural network model was trained and tested on the same dataset. The results are as follows:

- **Accuracy**: The neural network achieved an accuracy of 88.08%, which is lower than the Bayesian network's 97.07%. This suggests that, while the neural network is quite accurate, it may not be as reliable as the Bayesian model for this particular task.
- **Precision & Recall**:

Metric	No Stroke	Stroke
Precision	0.95 (95%)	0.84 (84%)
Recall	0.80 (80%)	0.96 (96%)

Table 6: Precision and Recall for the Neural Network Model

- **F1-score**: The F1-score for "No Stroke" predictions was 0.87 and for "Stroke" predictions was 0.89. Although these scores are high, they do not reach the harmonized level of precision and recall exhibited by the Bayesian network.
- **ROC Curve**: While commendable, the neural network's AUC of 0.93 is slightly below the Bayesian network's AUC of 0.98, indicating that the Bayesian network may be more adept at distinguishing between positive and negative classes in this context. For a visual representation, see Figure 4 in the Appendix.

The Bayesian network provides a transparent and probabilistic framework, which is particularly beneficial in medical settings where understanding the reasoning behind a prediction is as crucial as the prediction itself. On the other hand, neural networks operate as "black boxes," and despite their ability to handle large datasets and capture complex nonlinear relationships, they offer less interpretability. However, neural networks can be advantageous when dealing with unstructured data, such as images or text, where Bayesian networks may struggle without significant feature engineering.

5 Limitation and Possible Future Enhancement

5.1 Limitations

- **Comprehensiveness of Model Exploration**: Our investigation into Bayesian Networks was ambitious, yet the sheer number of potential structures makes it unfeasible to thoroughly explore each variant. While we endeavored to construct and test a wide range of network configurations, the reality of finite resources meant that some potentially valuable models may have gone unexamined.
- **Breadth of Comparative Analysis**: The ambition to conduct a broad comparative analysis of our Bayesian Network against a spectrum of alternative models was tempered by practical time constraints. Consequently, our comparison was limited to a single additional neural network model. This restricted scope means we may have missed opportunities to identify and analyze other advanced modeling techniques that could offer superior predictive performance or insights.
- **Comprehensiveness and Quality of Data**: The dataset used, while extensive and carefully curated, is not exhaustive. There are likely additional variables and complex interactions that influence the risk of stroke which our dataset does not capture. This limitation is inherent in any data-driven approach and underscores the need for continuous data enhancement and model refinement.
- **Generalizability of the Model**: Our model's predictions are as good as the data on which they are based. The specific characteristics of the dataset, reflective of a particular demographic or healthcare system, may not be universally applicable. This limitation highlights a potential risk when applying the model to different or broader populations without recalibration or consideration of local context.

- **Inherent Simplifications in Bayesian Modeling**: Bayesian networks, by their nature, apply conditional independence assumptions to model complex systems. These simplifications are necessary for computational tractability but may not always align with the nuanced reality of medical phenomena. As a result, some subtleties of stroke risk factors and their interrelations may not be fully represented in the model's structure.

5.2 Possible Future Enhancements

- **Broader Model Comparisons**: Expanding the scope to include a variety of modeling approaches, such as additional probabilistic models, advanced deep learning architectures, and ensemble strategies, may yield insights into more precise and reliable prediction methods.
- **Data Enrichment**: Enhancing the dataset with new variables informed by the latest medical research or by integrating additional datasets could provide a more nuanced understanding of stroke risk factors, leading to better predictions.
- **Dynamic Modeling**: Investigating dynamic Bayesian networks as a means to account for the temporal evolution of risk factors could capture how stroke risk changes over time with greater accuracy.
- **Mitigating Overfitting**: Applying regularization techniques could be crucial in maintaining model generalization, particularly as the complexity of the network increases with the addition of new variables and structures.
- **Interface Development**: The creation of a straightforward and intuitive interface would be a valuable tool for healthcare providers, enabling them to leverage the model for on-the-spot stroke risk assessments with ease.

6 Towards Responsible AI

6.1 Transparency

The Bayesian network at the heart of our project excels in transparency, offering a clear visual representation of the relationships between various health and lifestyle factors and their impact on stroke risk. This approach demystifies the predictive process, enabling healthcare professionals and patients alike to understand how and why certain conclusions are drawn. However, the interpretability of Bayesian networks can be limited by their complexity, particularly for those without a background in statistics or machine learning. To address this, we propose ongoing initiatives such as specialized training modules and interactive tools that simplify the interpretation of these networks. These resources would be designed to foster a broader understanding among healthcare practitioners, enhancing their ability to leverage the model's insights effectively.

6.2 Privacy

In the realm of healthcare, where sensitive personal data is involved, privacy is paramount. Our use of anonymized datasets serves as a foundational step in ensuring privacy. To further fortify privacy safeguards in clinical settings, we envision the implementation of stringent data governance policies, encompassing advanced anonymization techniques and robust access control mechanisms. There's an ever-present challenge of balancing data utility with privacy, especially considering the potential for re-identification in rich datasets. Addressing this requires not only technological solutions but also a strong ethical framework guiding data usage and privacy protection. Regular privacy audits and adherence to global data protection regulations are critical elements of this framework.

6.3 Fairness

Fairness in AI is a multifaceted challenge, especially in healthcare applications where biases in data can lead to unequal treatment outcomes. Our commitment to fairness involves conscientious data collection and model training practices, aimed at capturing a representative cross-section of the population. Despite these efforts, latent biases in healthcare data — a reflection of broader societal inequities — pose a persistent challenge. To combat this, we propose continuous bias monitoring mechanisms and iterative model training with updated, more equitable data. This approach necessitates a dynamic, rather than static, model development process, where fairness is an ongoing objective rather than a one-time achievement.

6.4 Safety and Robustness

In healthcare AI, where erroneous predictions can have serious implications, the safety and robustness of our Bayesian network are of utmost concern. We employ rigorous validation techniques and diverse evaluation metrics to ensure the model's accuracy and reliability. Recognizing that healthcare contexts and data patterns evolve, our model is designed for adaptability, with mechanisms for regular re-evaluation and updates. This proactive approach to model management helps mitigate risks associated with data drift and changing healthcare landscapes. Additionally, establishing clear guidelines for handling ambiguous or uncertain cases is crucial, ensuring that the model operates within its limits of reliability.

6.5 Human Factors

The role of our Bayesian model is to support, not supplant, the expertise of healthcare professionals. It serves as a decision-support tool, providing insights based on data-driven predictions. The primary risk here is the potential for automation bias, where clinicians might overly rely on the model's output. To counter this, we stress the importance of human judgment and contextual knowledge in interpreting model predictions. Educational programs and decision-making frameworks that integrate AI insights with clinical expertise are proposed to enhance the symbiosis between the model and human decision-makers.

6.6 Governance and Regulation

As AI technologies increasingly permeate healthcare, robust governance and adherence to regulatory standards become crucial. Our project, while academic, lays the groundwork for responsible AI deployment in clinical settings. Real-world application, however, calls for compliance with a complex and evolving regulatory landscape. This challenge necessitates a commitment to continuous learning and adaptation, ensuring that the model's use is not only effective but also ethical and compliant. Regular ethical audits, transparency in model development and deployment, and a steadfast commitment to upholding AI ethics principles are fundamental to navigating this landscape responsibly.

In summary, our project demonstrates a conscientious approach to developing AI in healthcare, aligning with the principles of responsible AI. While there are inherent challenges in each of these domains, our commitment to ongoing improvement, ethical considerations, and adaptability positions our model as a valuable tool in the landscape of healthcare AI.

7 Conclusion

In this study, we delved into the potential of Bayesian Networks to predict stroke risk, harnessing both prior knowledge and observed data. The results showcased the efficacy of our chosen model, underpinned by both data-driven insights and established medical literature. While the model exhibited promising accuracy and robustness, we also identified areas for improvement and potential future enhancements. The intersection of artificial intelligence and healthcare holds immense promise, and our work underscores the pivotal role Bayesian modeling can play in advancing predictive healthcare. As we move forward, refining and expanding upon this foundation could pave the way for more nuanced and reliable medical predictions, ultimately aiding in better patient outcomes and preventive healthcare strategies.

References

[1] Stroke prediction models: A systematic review https://www.ijser.org/researchpaper/Stroke-Prediction-Models-A-Systematic-Review.pdf

[2] Stroke risk prediction with machine learning techniques https://www.mdpi.com/1424-8220/22/13/4670

[3] Early prediction of ischemic stroke using machine learning boosting (2022), https://ieeexplore.ieee.org/document/10205861/

[4] Association, A.H.: How high blood pressure can lead to stroke (2022), https://www.heart.org/en/health-topics/high-blood-pressure/health-threats-from-high-blood-pressure/how-high-blood-pressure-can-lead-to-stroke

[5] Association, A.H.: Smoking, high blood pressure and your health - american heart association (2023), https://www.heart.org/en/health-topics/high-blood-pressure/changes-you-can-make-to-manage-high-blood-pressure/smoking-high-blood-pressure-and-your-health

[6] CDC: Heart disease and stroke — smoking and tobacco use — cdc (2023), https://www.cdc.gov/tobacco/basic_information/health_effects/heart_disease/index.htm

[7] Clinic, M.: Hypertensive crisis: What are the symptoms? (2022), https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/expert-answers/hypertensive-crisis/faq-20058491

[8] Fedesoriano: Stroke prediction dataset. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (2021)

[9] GitHub: Exploratory data analysis of stroke dataset in r (2023), https://github.com/djthorne333/Exploratory-Data-Analysis-of-Stroke-Dataset-in-R

[10] Health, B.P.: Long-term body mass index changes in overweight and obese adults and the risk of heart failure, cardiovascular disease and mortality: a cohort study of over 260,000 adults in the uk (2021), https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-021-10606-1

[11] in Health, L., Disease: Body mass index is associated with blood pressure and vital capacity in medical students (2023), https://lipidworld.biomedcentral.com/articles/10.1186/s12944-023-01920-1

[12] Health, P.D.: Ideal algorithms in healthcare: Explainable, dynamic, precise ... (2022), https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000006

[13] Healthline: What is the risk of having a stroke by age range? (2022), https://www.healthline.com/health/age-range-for-stroke

[14] Healthline: How are hypertension, heart disease, and stroke related? (2023), https://www.healthline.com/health/high-blood-pressure-hypertension/how-are-hypertension-heart-disease-and-stroke-related

[15] Kaur, M., et al.: Early stroke prediction methods for prevention of strokes (2022), https://www.hindawi.com/journals/bn/2022/7725597/

[16] Li, F., Ding, P., Mealli, F.: Bayesian causal inference: A critical review. arXiv preprint arXiv:2206.15460 (2022), https://doi.org/10.48550/arXiv.2206.15460

[17] Library, I.X.D.: Heart disease prediction using stacking model with balancing techniques ... (2022), https://ieeexplore.ieee.org/document/10287281

[18] Mainali, S., Darsie, M.E., Smetana, K.S.: Machine learning in action: Stroke diagnosis and outcome prediction. Frontiers in Neurology **12** (2021), https://www.frontiersin.org/articles/10.3389/fneur.2021.734345/full

[19] Mainali, S., Darsie, M.E., Smetana, K.S.: Machine learning in action: Stroke diagnosis and outcome prediction. Frontiers in Neurology **12** (2021), https://www.frontiersin.org/articles/10.3389/fneur.2021.734345/full

[20] Medicine, J.H.: Smoking and cardiovascular disease — johns hopkins medicine (2023), https://www.hopkinsmedicine.org/health/conditions-and-diseases/smoking-and-cardiovascular-disease

[21] medRxiv: From tabulated data to knowledge graph: A novel way of improving the performance of the classification models in the healthcare data (2021), https://www.medrxiv.org/content/medrxiv/early/2021/06/12/2021.06.09.21258123.full.pdf

[22] Network, J.: Age-adjusted mortality rates and age and risk–associated contributions to change in heart disease and stroke mortality (2023), https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2790434

[23] NIH: Scientists investigate link between smoking and high blood pressure — nhlbi, nih (2021), https://www.nhlbi.nih.gov/news/2021/scientists-investigate-link-between-smoking-and-high-blood-pressure

[24] Organization, W.H.: World stroke day (2021), https://www.who.int/southeastasia/news/detail/28-10-2021-world-stroke-day

[25] Paolucci, I., et al.: Bayesian parametric models for survival prediction in medical applications (2023), https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-023-02059-4

[26] Research, C.: Smoking and gender (2002), https://academic.oup.com/cardiovascres/article/53/3/568/325628

[27] Today, M.N.: Can smoking cause a stroke? risk, statistics, and more (2023), https://www.medicalnewstoday.com/articles/can-smoking-cause-a-stroke

[28] Vidhya, A.: How to create a stroke prediction model? (2021), https://www.analyticsvidhya.com/blog/2021/05/how-to-create-a-stroke-prediction-model/

[29] Wang, W., Kiik, M., Peek, N., Curcin, V., Marshall, I.J., Rudd, A.G., Wang, Y., Douiri, A., Wolfe, C.D., Bray, B.: A systematic review of machine learning models for predicting outcomes of stroke with structured data. PLOS ONE **15**(6) (2020), https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0234722

[30] Wang, W., Kiik, M., Peek, N., Curcin, V., Marshall, I.J., Rudd, A.G., Wang, Y., Douiri, A., Wolfe, C.D.A., Bray, B.: A systematic review of machine learning models for predicting outcomes of stroke with structured data. PLOS ONE **15**(6) (2020), https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0234722

Appendix: Diagrams for Alternative Models

Neural Network Model Analysis

Visual Representation

The following diagram represents the neural network architecture used in the alternative deep learning model. It consists of an input layer with 11 nodes, two hidden layers with 128 and 64 nodes respectively, and an output layer with a single node.

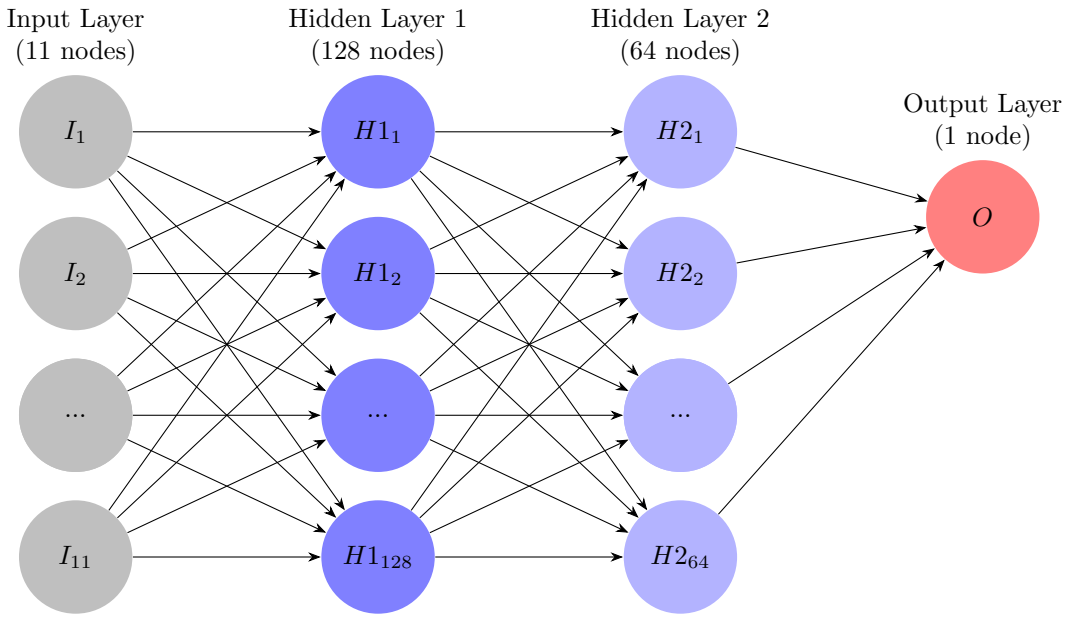

Figure 3: Neural Network Model Architecture

ROC Curve for the Neural Network Model

Below is the ROC curve for the neural network model.

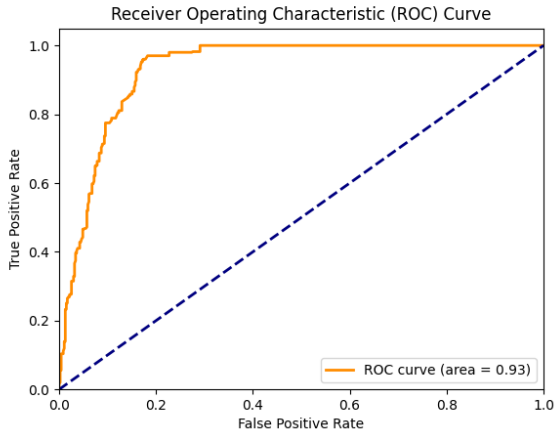

Figure 4: ROC Curve for More Complex Model

Simpler Model Analysis

Visual Representation

This Diagram is the simplified Bayesian network for stroke prediction:

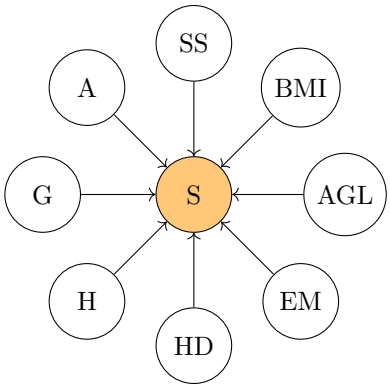

Figure 5: Simplified Bayesian network for stroke prediction.

Legend:

Symbol	Description	Symbol	Description	Symbol	Description
A	age	G	gender	H	hypertension
HD	heart_disease	S	stroke	EM	ever_married
SS	smoking_status	AGL	avg_glucose_level	BMI	BMI

Table 7: Legend for the Symbols

ROC Curve for the Simpler Model This diagram depicts the Receiver Operating Characteristic (ROC) curve for the simpler model, illustrating its performance in discriminating between the positive and negative classes.

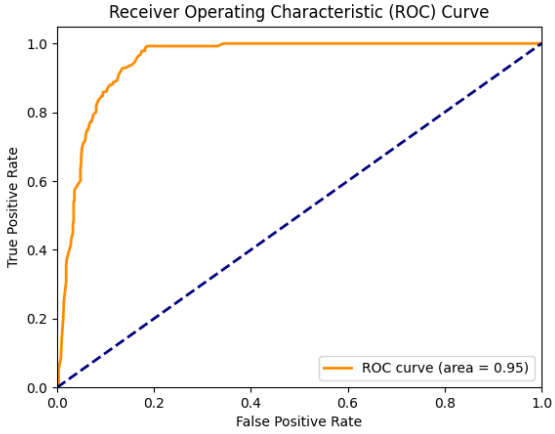

Figure 6: ROC Curve for Simpler Model

Complex Model Analysis

Visual Representation

Here is the complex Bayesian network for stroke prediction:

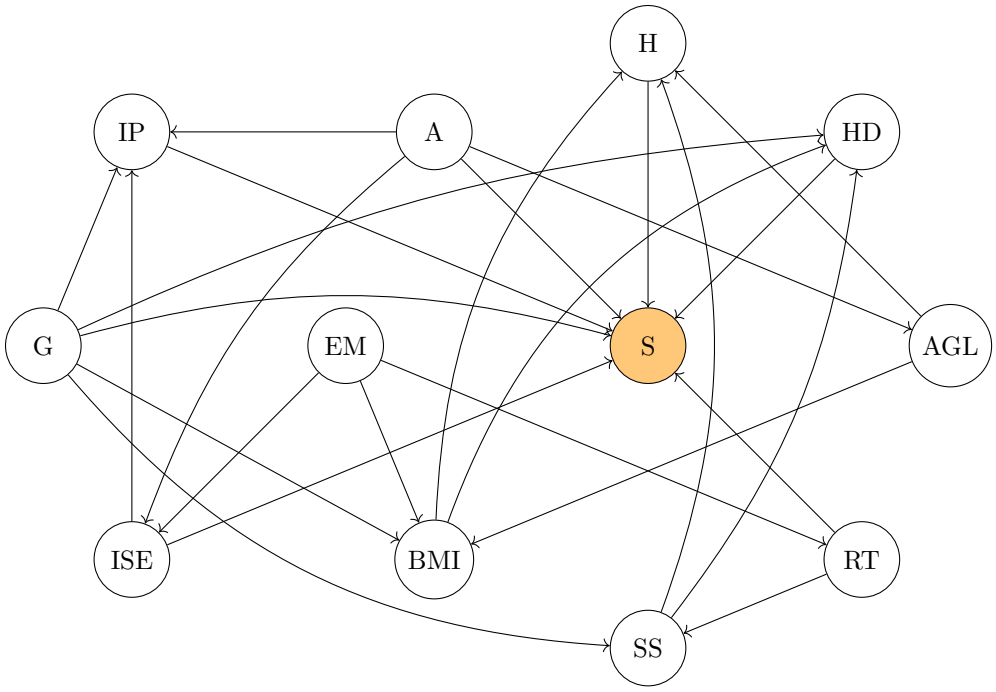

Figure 7: More Complex Bayesian Network for Stroke Prediction.

Legend:

Symbol	Description	Symbol	Description	Symbol	Description
A	age	G	gender	H	hypertension
HD	heart_disease	S	stroke	EM	ever_married
RT	residence_type	AGL	avg_glucose_level	BMI	BMI
SS	smoking_status	IP	is_private	ISE	is_self_employed

Table 8: Legend for the Symbols

ROC Curve for the More Complex Model Below is the ROC curve for the more complex model. Despite its detailed architecture, its performance is compared to the simpler and the proposed models.

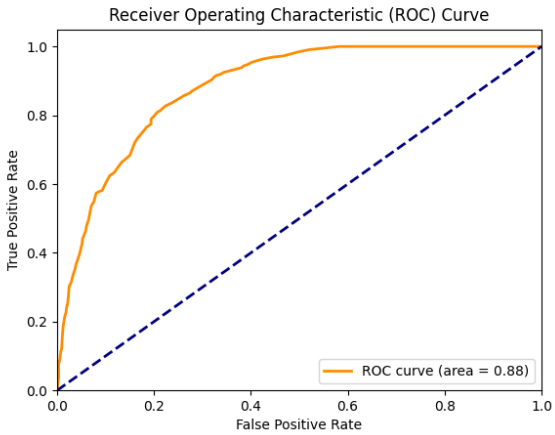

Figure 8: ROC Curve for More Complex Model