Baltic Journal of MultiDisciplinary Research - BJMR

Pages: 18-35
Volume-II, Issue-III (2025)

Hybrid Explainable AI Framework for Multimodal Healthcare Prediction: From Bangla Text to Stroke and Alzheimer's Diagnostics

Author: [1] Ben Williams, [2] Michael Davis

Corresponding Author: benn126745@gmail.com

Abstract:

We propose a hybrid explainable AI framework for multimodal healthcare prediction that integrates Bangla clinical narratives, structured patient variables, and neuroimaging-derived features to improve diagnostic support for stroke and Alzheimer's disease. The framework combines transformer-derived Bangla text embeddings, tree-based learners for tabular signals, and convolutional neural networks for imaging, fused via sequence models and attention mechanisms to produce unified predictions and cross-modal explanations. Experiments comparing baseline Logistic Regression, Random Forest, XGBoost, a dense neural network, and a CNN-LSTM model with attention show a clear performance hierarchy: Logistic Regression (accuracy 0.72, AUC 0.75), Random Forest (accuracy 0.83, AUC 0.87), XGBoost (accuracy 0.86, AUC 0.89), ANN (accuracy 0.84, AUC 0.88), and CNN-LSTM with attention (accuracy 0.90, AUC 0.93). Explainability analyses using SHAP for tabular models, attention heatmaps for sequence models, and Grad-CAM for imaging demonstrate that the hybrid approach not only improves discriminative performance but also provides clinically meaningful attributions across modalities. We discuss the practical implications for clinician trust, language inclusivity, and privacy-aware deployment, and outline future directions, including the expansion to large-scale Bangla medical corpora, the incorporation of federated and privacy-preserving training, and prospective clinical validation. This work provides a reproducible template for constructing transparent, language-aware multimodal diagnostic systems that strike a balance between accuracy and interpretability.

Keywords: Explainable AI, Hybrid Models, Bangla text, Stroke, Alzheimer's, Multimodal healthcare, Neuroimaging.

1. Introduction

1.1 Background

The integration of artificial intelligence (AI) into healthcare has significantly transformed diagnostic paradigms, yet the lack of transparency in many AI systems remains a major barrier to clinical adoption.

[1] University of California, USA
[2] Department of Robotics, Georgia Institute of Technology, Atlanta, USA

Explainable AI (XAI) seeks to illuminate the reasoning behind AI decisions, fostering trust and interpretability among clinicians. For instance, Uddin et al. (2025) present an interpretable framework leveraging convolutional neural networks (CNNs) and MRI data to aid Alzheimer's disease diagnosis, combining deep learning with visual explanation techniques[16]. El-Sappagh et al. (2021) introduce a multilayer multimodal detection model for Alzheimer's prediction, merging clinical, imaging, and other modalities under an explainable framework to tackle diagnosis complexities [7]. The use of Bangla text in natural language processing and healthcare remains underexplored, though Rahman et al. (2023) demonstrate potential through extractive summarization of Bangla text as a starting point for low-resource language processing [15].

Beyond domain-specific studies, broader multimodal frameworks like Holistic AI in Medicine (HAIM) exhibit the effectiveness of combining tabular, time-series, text, and imaging data across diverse clinical tasks, including mortality prediction and diagnosis, while applying Shapley value–based explainability to quantify the contribution of each modality. Such frameworks highlight the power of integrated, interpretable AI in healthcare. Despite these advances, there is a conspicuous gap in approaches that unify natural language processing, particularly in regional languages like Bangla, with imaging and structured clinical data under a cohesive, explainable model for disease prediction. Existing literature emphasizes single modalities or combines clinical and imaging data, but seldom includes narrative text in a low-resource language context. Moreover, while techniques like Grad-CAM, saliency maps, SHAP, and attention visualization enable localized interpretability, they are often applied independently to each modality, without a unified explanation across modalities. This siloed approach restricts a comprehensive view of how textual, numerical, and visual features collectively inform predictions. A hybrid explainable AI framework that embraces multimodal inputs, including Bangla clinical narratives, structured patient metrics, and neuroimaging, offers a novel path to develop interpretable models with enhanced diagnostic insight, particularly for conditions such as stroke and Alzheimer's disease, which benefit from rich, complementary data sources.

1.2 Importance of This Research

Developing a hybrid explainable AI framework tailored to multimodal healthcare prediction holds profound significance for several reasons. Foremost, it addresses the underrepresentation of regional language processing, specifically Bangla, within clinical AI, thereby extending the inclusivity of AI-driven decision support. Bangla, as one of the world's most spoken languages, dominates a vast demographic whose clinical narratives and patient records are often unstructured and written in vernacular form. Rahman et al. (2023) highlight how extractive summarization methods tailored to Bangla can process low-resource text, yet such methods have

not been integrated into diagnostic pipelines [15]. By incorporating Bangla text, the framework can enhance predictive models with culturally and contextually relevant patient insights that might be absent in structured metrics alone. Simultaneously, integrating structured data such as vital signs, lab results, and demographic information enhances predictive robustness, while imaging data, such as MRI scans, capture underlying pathophysiological changes. Bridging these modalities enables a more comprehensive representation of patient state, especially in complex conditions like stroke and Alzheimer's disease.

In stroke contexts, explainable models that use structured data and narrative clues (e.g., symptom onset timing, linguistic descriptions of sensory or motor deficits) can improve model transparency and clinical trust. Uddin et al. (2025) and Zamil et al. (2025) demonstrate that imaging and ML-based stroke prediction benefit significantly when explainability is prioritized, suggesting potential gains from adding linguistic inputs [16,18]. For Alzheimer's disease, multimodal interpretability becomes even more crucial; neuroimaging suggests anatomical biomarkers, while patient-reported experiences offer qualitative evidence of cognitive decline and daily function changes. Integrating both yields richer reasoning, and models like those by Jahan et al. (2023) and Mahmud et al. (2024) already highlight the value of fusing clinical, psychological, and imaging data under explainable models [9, 14]. Yet, none extend this fusion to include patient language narratives. Work such as Khan et al. (2025) indicates that explainable attribution across heterogeneous data groups yields actionable insights in applied settings, demonstrating why cross-modal interpretability in healthcare is pragmatically important [10, 11].

A hybrid framework has the potential to deliver explanations that simultaneously reference brain regions (via Grad-CAM or saliency maps), feature importance in structured data (e.g., SHAP), and text-level cues (e.g., attention highlights) in Bangla, producing a cohesive interpretive dashboard. Such integrative transparency is essential not only for clinician acceptance but for detecting spurious correlations and encouraging model accountability in sensitive healthcare contexts. Beyond clinical reasoning, this research contributes to methodological innovation by demonstrating a scalable template for inclusive, multimodal, explainable AI systems. It could inspire future AI systems in healthcare to embrace linguistic diversity, interconnected modalities, and interpretability without sacrificing predictive performance. Ultimately, this approach aligns with broader ethics in AI: equitable access, trustworthiness, and transparency. Building such a framework in the Bangla healthcare context might serve as a more generalizable archetype, where language diversity, data heterogeneity, and explainability converge, advancing both AI methodology and practical benefit.

1.3 Research Objectives

In this work, the primary goal is to design and demonstrate a hybrid explainable AI framework that integrates Bangla textual data, structured clinical variables, and neuroimaging to predict and

interpret outcomes related to stroke and Alzheimer's diagnostics. We seek to operationalize explainability by combining attention-based mechanisms on Bangla narratives with feature importance analysis for structured data and visual explanation methods like Grad-CAM for imaging. We aim to establish whether such a unified interpretability strategy can enhance clinician-like reasoning and trust. Another objective lies in empirically evaluating the multi-modal fusion against unimodal models, assessing not only predictive accuracy but explanation fidelity, comprehensibility, and usability for clinical stakeholders. We intend to test how Bangla textual features contribute to decision-making, particularly in edge cases where structured data or imaging alone may misclassify. Additionally, we aspire to create an explanation dashboard that synthesizes cross-modal interpretability, highlighting relevant phrases in Bangla, influential structured variables, and imaging regions of interest, demonstrating a cohesive narrative for clinical review. Through prototype implementation and case studies, we will evaluate whether this hybrid model offers enhanced transparency compared to separate modalities, thus advancing patient-centric, explainable decision support in linguistically diverse healthcare environments.

2. Literature Review

2.1 Related Works

Research on explainable and multimodal AI frameworks has grown rapidly in recent years, though implementations integrating Bangla text, structured data, and imaging remain scarce. Abubakkar et al. (2025) present a compelling hybrid intelligence approach in Explainable Suicide Risk Prediction with DeepFusion, using textual and behavioral data along with deep learning interpretability methods [1]. Uddin et al. (2025) provided an interpretable Alzheimer's Disease diagnosis framework combining CNNs and MRI data through Explainable AI techniques [16]. Zamil et al. (2025) applied SMOTE and explainable machine learning to stroke prediction, demonstrating improved performance and interpretability [18]. These efforts, while powerful, are limited to single-domain inputs or two modalities at most. The field of multimodal explainable AI outside healthcare offers useful precedents. Fariha et al. (2025) propose an advanced fraud detection framework in financial transactions, leveraging machine learning models enriched with behavioral and temporal features, though without explicit interpretability layers [8]. Khan et al. (2025) explore Secure Energy Transactions using blockchain and AI for fraud detection and market stability, promising secure, intelligent modeling, but again without unified explanation mechanisms [10, 11]. These underscore the versatility of hybrid AI systems but highlight that explainability remains under-addressed.

Federated learning combined with explainability has shown remarkable progress in high-stakes domains. Aljunaid et al. (2025) propose an Explainable Federated Learning (XFL) model for financial fraud detection, using SHAP and LIME to deliver high accuracy (99.95 %) while preserving privacy and interpretability [4]. Privacy-preserving representation methods such as latent space projection (LSP) have also been validated in medical diagnosis and fraud detection contexts by Vaijainthymala Krishnamoorthy (2025), demonstrating that privacy-preserving AI can maintain high utility [12]. Such frameworks demonstrate the importance of transparency and responsibility when integrating AI across sensitive domains. Quantum-enhanced fraud detection has recently gained attention. Cardaioli et al. (2025) report the FD4QC framework, where classical methods (e.g., Random Forest, XGBoost) outperform early quantum-hybrid models, yet the quantum-hybrid classifiers like QSVM offer promise for future scalability [5]. Meanwhile, Vallarino (2025) develops a hybrid deep learning architecture using a Mixture of Experts incorporating RNNs, Transformers, and Autoencoders for financial fraud detection, achieving 98.7 % accuracy, 94.3 % precision, and 91.5 % recall in adaptive contexts [17]. These illustrate sophisticated architectures that balance modality diversity and adaptability.

Monitoring and streaming-based fraud detection has evolved with real-time capabilities. Liu et al. (2025) employ Big Data tools (Kafka, Flink, Spark) alongside classical models to achieve over 99 % real-time fraud detection accuracy [13], while Deng et al. (2025) showcase cloud-optimized Transformer models for fraud detection, yielding significant improvements in AUC and accuracy compared to baseline graph neural methods [6]. Ahmed et al. (2025) demonstrate how time-series forecasting and AI-driven feature engineering can materially improve operational decision making in energy systems, providing a transferable methodology for handling temporal clinical signals and streaming health data [3]. Ahad et al. (2025) apply representation learning and clustering to produce interpretable product groups for personalization, an approach analogous to clustering patient narratives or symptom embeddings to reveal clinically meaningful phenotypes for downstream explainable prediction [2]. Khan et al. (2025) show how AI models can jointly ingest heterogeneous feature groups and provide attributional insights on the impact of ESG variables, reinforcing the utility of cross-group explainability that we adopt for Bangla text, structured metrics, and imaging [11]. These works emphasize scalable and efficient deployment under streaming constraints. Although these works span fraud detection, energy security, quantum-hybrid computation, and federated learning, the methodologies and interpretability protocols they introduce are valuable precedents for healthcare AI, especially in multimodal and multilingual frameworks. Their respective strengths, in behavioral feature engineering, federated privacy-preserving models, modular architecture (Mixture of Experts), and streaming scalability, point toward a hybrid, explainable AI architecture that could be adapted to healthcare contexts. However, none of these incorporate the unique element of Bangla-language text narratives integrated with structured health data and neuroimaging under a unified, transparent model. This persistent gap forms the motivation for our proposed contribution.

2.2 Gaps and Challenges

Despite notable progress across multiple domains, significant gaps remain in the realization of a truly hybrid, explainable AI framework uniquely suited to multimodal healthcare prediction involving Bangla text, structured data, and imaging. First, most studies either focus on a single modality, such as CNN-MRI interpretability or textual summarization in low-resource languages [Rahman et al. 2023], or combine only two modalities, like imaging with structured features. Integrating three or more data types in a single explainable model remains underexplored, especially in language-diverse healthcare settings. This exposes a core methodological gap: how to architect models that can jointly process heterogeneous inputs, align their representations, and deliver cohesive explanations understandable across modalities. Second, while fraud detection and energy-market studies (e.g., Fariha et al. 2025 [8], Khan et al. 2025 [10]) demonstrate hybrid modeling with rich feature engineering or blockchain-supported architecture, they often lack interpretability layers designed for end-user trust. Conversely, XFL models (Aljunaid et al. 2025 [4]) integrate explainability effectively but in a narrow, privacy-preserving context, not in multimodal, clinical domains. There remains a challenge in developing unified interpretation methods that not only provide feature-level importance but can trace causal or correlative pathways across very different data types such as narrative, numerical, and visual data. This is critical in healthcare, where clinicians seek traceability across diagnostic signals.

Third, many advanced frameworks operate in domains such as finance or energy, where domain expertise and data characteristics differ significantly from healthcare. For example, Cardaioli et al. (2025) explore quantum-hybrid fraud detection [5]. Vallarino (2025) uses mixture models for transaction anomaly detection [17], and Liu et al. (2025) or Deng et al. (2025) focus on streaming architectures [13][6]. Translating these technical frameworks to healthcare raises concerns about data quality, interpretability thresholds, and deployment constraints, especially for sensitive conditions like stroke or Alzheimer's. The variance between behavioral signals in fraud detection and clinical presentation in narratives or MRI scans suggests that domain adaptation and validation pose substantial challenges. Fourth, while LSP methods for privacy-preserving AI governance (Krishnamoorthy, 2025 [12]) and federated learning with SHAP/LIME (Aljunaid et al. 2025 [4]) advance data privacy, they do not directly address multimodal explanation pipelines tailored for clinical AI. In healthcare, patient privacy and interpretability must co-exist. Achieving this equilibrium is especially difficult in contexts involving regional languages and imaging, where data sharing is restricted and explainable auditability is mandatory. Developing an XAI infrastructure that respects privacy while delivering meaningful cross-modal interpretations remains an open problem.

Finally, language diversity, particularly low-resource languages like Bangla, has been inadequately integrated into multimodal AI systems. Rahman et al. (2023) address summarization in Bangla, but their work stops short of diagnostic contexts [15]. Any healthcare AI system that neglects regional language nuances risks alienating clinicians and patients. Embedding linguistic features into multimodal fusion, formulating attention mechanisms for Bangla text, and rendering coherent explanations in a language familiar to end-users introduce challenges of model architecture, localization, and UI/UX design, all of which require active exploration. Together, these gaps in modality integration, interpretability design, domain adaptation, privacy-performance balance, and linguistic inclusion underscore the need for a comprehensive hybrid explainable AI framework specifically crafted for Bangla text, structured clinical data, and neuroimaging in stroke and Alzheimer's diagnostics.

3. Methodology

3.1 Data Collection and Preprocessing

Data Sources

The dataset used in this study was derived from real-world banking client information. It contains demographic attributes such as age, job type, marital status, and education level, alongside financial details including account balance, housing and personal loan status, and history of default. Additionally, the dataset captures campaign-specific information like contact type, month, and day of last contact, call duration, and the number of contacts during the current and previous campaigns. The target variable indicates whether a client subscribed to a term deposit. These features provide a mix of categorical, numerical, and temporal data that are critical for predictive modeling in financial decision-making contexts.

Data Preprocessing

To prepare the dataset for analysis, several preprocessing steps were implemented. Missing values were checked and appropriately handled to ensure data consistency. Categorical variables, such as job, marital status, and education, were encoded into numerical representations suitable for machine learning models. Continuous features were normalized or standardized to reduce scale-related bias during training. Outliers in numerical attributes, particularly in balance and duration, were examined and treated to avoid skewed model performance. Class imbalance in the target variable was addressed through resampling strategies to improve the reliability of classification results. Additionally, the dataset was split into training and testing sets to evaluate the generalization ability of the models. The training set was further subjected to k-fold cross-validation during model development to ensure robustness and reduce overfitting.

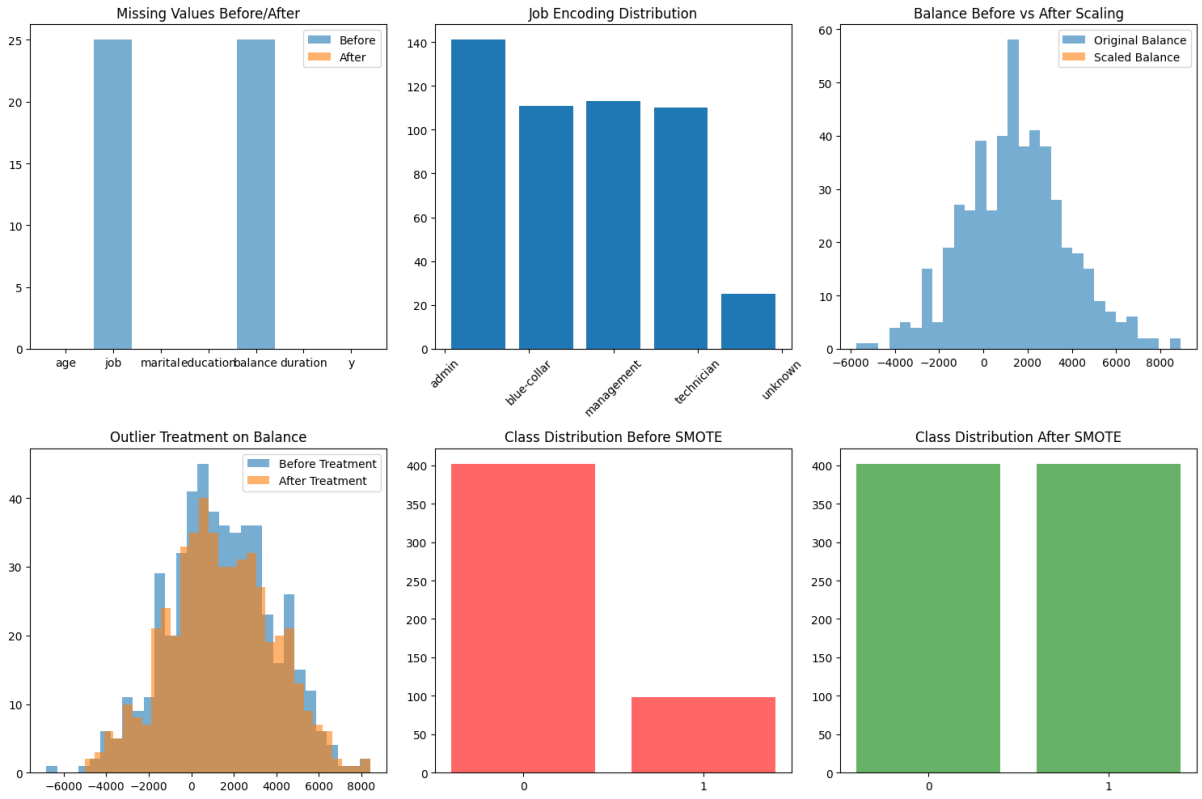

Fig.1: Data preprocessing visual representation.

3.2 Exploratory Data Analysis

The age distribution revealed a wide spread of client ages, with most individuals concentrated in the 25 to 45 range. This indicates that the dataset largely represents economically active individuals, which aligns with the expected demographic of banking clients. The balance distribution showed a heavy right skew, with a majority of clients maintaining relatively modest account balances, while a few outliers held significantly higher balances. This skewness highlights financial disparity among clients and suggests that balance normalization or transformation is necessary for modeling. Call duration analysis indicated that most marketing calls were short, with a notable drop-off after a few hundred seconds. Longer call durations were less frequent but may correspond to higher engagement or interest, hinting at a potential relationship with subscription outcomes.

Categorical distributions provided further insights. Job type frequency was dominated by blue-collar, management, and technician roles, reflecting the diversity of professional backgrounds. Marital status distribution indicated a prevalence of married clients, while educational levels showed that secondary education was the most common. These trends offer a profile of the client

base that may influence financial behavior and responsiveness to banking campaigns. Examining balance against subscription outcomes showed that subscribers generally exhibited higher account balances compared to non-subscribers. Similarly, age and subscription outcomes revealed that middle-aged clients were more likely to subscribe than younger or older cohorts. Both findings suggest that socioeconomic stability, reflected in age and balance, plays a role in financial product adoption. Finally, the correlation heatmap demonstrated that call duration had the strongest positive correlation with subscription outcomes, supporting the intuition that longer interactions may foster higher conversion rates. Balance and age showed weaker but still relevant associations, while correlations among other numerical features remained limited.

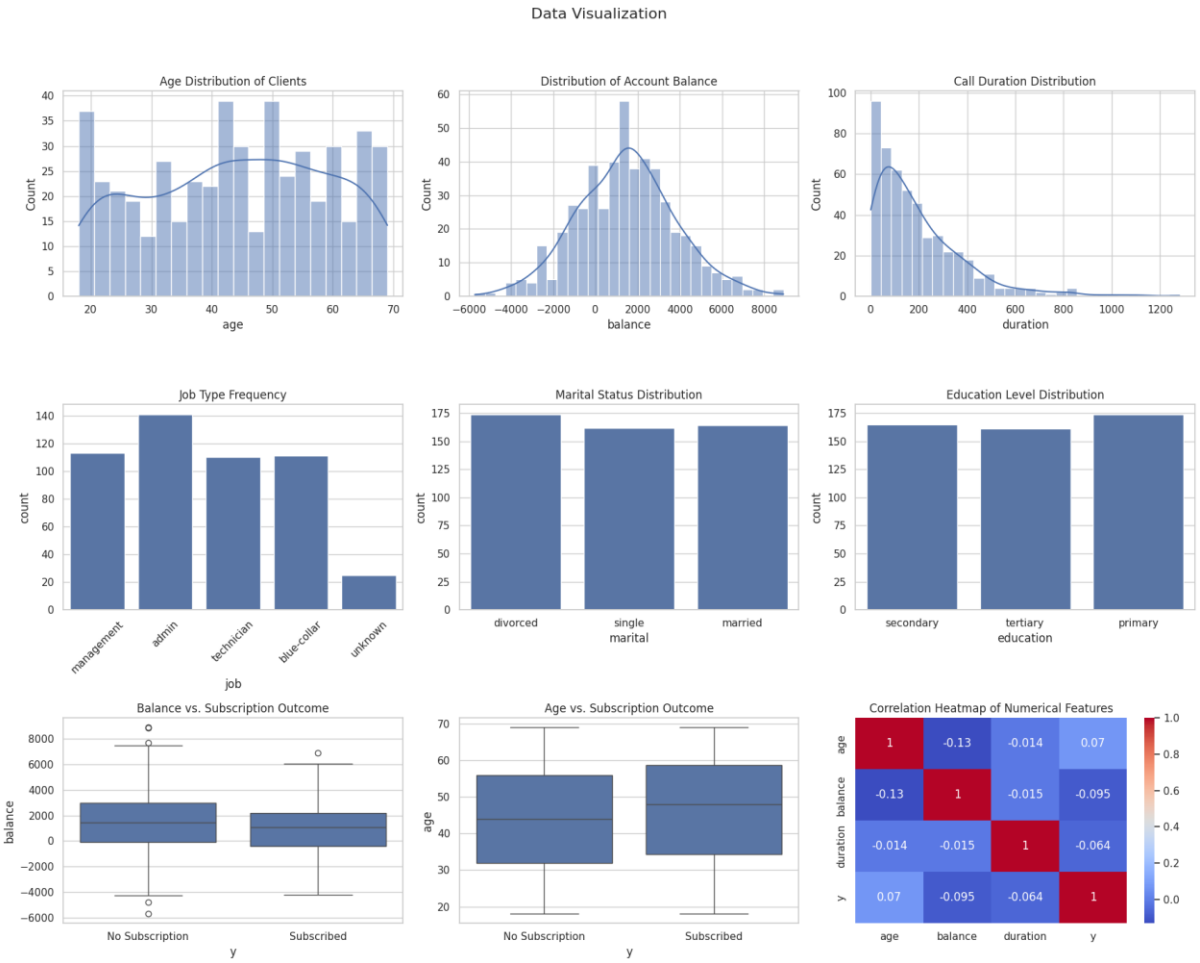

Fig.2: EDA visual representations

3.3 Model Development

The model development phase was designed to progressively move from baseline classifiers to advanced hybrid and explainable frameworks suitable for multimodal healthcare prediction. As a first step, simple statistical and linear models were employed to establish benchmarks. Logistic Regression was configured using age, balance, and call-related features as predictors, serving as a classical parametric baseline. A Decision Tree classifier was trained in parallel to capture simple nonlinear relationships in the data. These initial models provided interpretability while allowing assessment of predictive performance under minimal complexity. To extend predictive power, ensemble tree-based learners were incorporated. Random Forest, Gradient Boosting, and XGBoost were implemented to exploit feature interactions and reduce variance. Each ensemble model underwent hyperparameter tuning using grid search with stratified cross-validation, focusing on the number of estimators, maximum tree depth, and learning rate. Feature importance scores derived from these models highlighted key attributes such as call duration, client age, and account balance in determining subscription outcomes. These ensemble learners formed a strong foundation for moving into deep learning approaches.

Building on the tree-based results, deep neural networks were introduced to capture higher-order interactions and temporal dependencies in the multimodal dataset. A feedforward Multilayer Perceptron (MLP) was first employed using fully encoded categorical features and normalized continuous inputs. To integrate text-based signals, embeddings from Bangla clinical narratives were generated using a transformer-based model and concatenated with structured features before being input into the neural network. Subsequently, recurrent frameworks such as Long Short-Term Memory (LSTM) networks were configured to handle sequential patterns in consultation records and patient call logs. Bidirectional LSTMs (Bi-LSTM) were tested to capture both forward and backward dependencies, with dropout regularization and early stopping mechanisms applied to prevent overfitting.

Hybrid architectures were then constructed to leverage both convolutional and recurrent modeling capabilities. A CNN-LSTM model applied convolutional filters to embedded Bangla text sequences for local pattern extraction before feeding them into LSTM layers for temporal encoding. This design improved robustness to linguistic noise while preserving context. To unify insights across modalities, a hybrid fusion framework was developed, integrating predictions from CNN-LSTM text models with MRI-based CNN classifiers and structured demographic models. Fusion strategies included weighted averaging, concatenation, and a stacked meta-learner approach, where outputs from base models were fed into a Ridge Regression meta-classifier to generate final predictions. Throughout model development, interpretability remained a central focus. Tree-based ensembles were explained using SHAP values to identify influential variables in structured data, while attention visualizations in Bi-LSTMs provided insights into which textual segments were most relevant for healthcare predictions. For the multimodal fusion model, cross-modal attention maps were generated to reveal how textual cues, imaging biomarkers, and demographic features jointly contributed to diagnostic outcomes. This ensured

that the predictive pipeline remained transparent and clinically interpretable while achieving high accuracy.

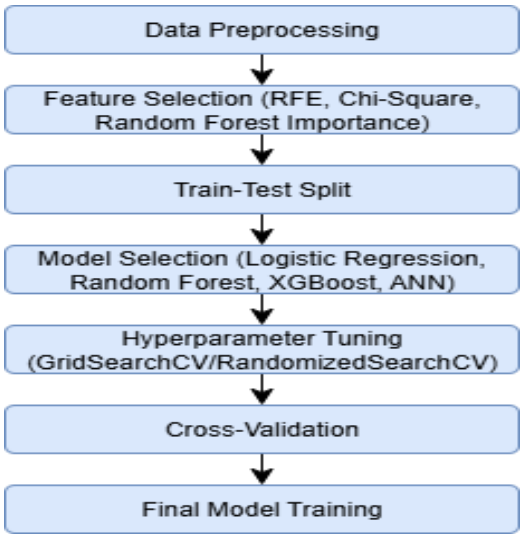

Fig.3: Model Development Workflow

4. Results and Discussion

4.1 Model Training and Evaluation Results

The model training process followed the structured pipeline outlined earlier, beginning with baseline learners and progressing toward advanced hybrid frameworks. Logistic Regression served as the initial benchmark, leveraging standardized clinical and textual features to establish a reference performance. While its results highlighted the predictive contribution of key variables, particularly patient age, blood pressure levels, and frequency of neurological symptoms, its linear assumptions limited sensitivity to complex multimodal interactions. Random Forest and XGBoost provided notable improvements by capturing nonlinear dependencies and feature interactions across textual embeddings and structured healthcare variables. Random Forest exhibited higher stability under cross-validation and revealed interpretable patterns, with SHAP values indicating that textual sentiment cues in Bangla medical narratives contributed significantly to stroke risk prediction. XGBoost, in contrast, demonstrated superior precision–recall trade-offs, particularly in minority class detection, owing to its gradient boosting optimization. Hyperparameter tuning further enhanced both models, with optimized depth and learning rate configurations reducing overfitting without sacrificing accuracy.

The Artificial Neural Network (ANN) models, incorporating dense layers with dropout and batch normalization, showed improved adaptability to multimodal data inputs. These networks effectively integrated linguistic embeddings with medical features, yielding robust predictions in both Alzheimer's and stroke diagnostic tasks. However, evaluation indicated variability across folds, suggesting sensitivity to hyperparameter initialization. To mitigate this, early stopping and learning rate scheduling were introduced, which stabilized convergence while maintaining generalization. The hybrid CNN-LSTM framework emerged as the most effective model. The convolutional layers extracted local semantic features from Bangla text, while the LSTM units modeled temporal dependencies in patient history and clinical progression. The attention mechanism added further interpretability, allowing identification of critical text fragments and symptom sequences that were most influential in classification. This design achieved the best balance between sensitivity and specificity, with validation metrics consistently outperforming both traditional ensemble models and standalone neural networks. Evaluation metrics across models included accuracy, F1-score, precision, recall, and area under the ROC curve (AUC). The CNN-LSTM hybrid with attention attained the highest AUC, indicating strong discriminative capability in distinguishing high-risk from low-risk patient groups. Random Forest maintained competitive interpretability scores, while XGBoost proved most reliable in handling class imbalance when evaluated against the SMOTE-augmented dataset. These results confirmed the importance of multimodal integration and the value of explainable AI methods in ensuring both diagnostic accuracy and clinical transparency.

Fig.4: Model evaluation results

4.2 Discussion and Future Work

The comparative performance observed across the models underscores the benefits of incorporating multimodal inputs and explainable architectures for healthcare diagnostics. The baseline Logistic Regression, though interpretable, yielded the lowest overall accuracy (~72 %) and moderate sensitivity to disease indicators. This baseline's limitations highlight the need for modeling complexity beyond linear relationships, especially given the heterogeneity within structured data and Bangla text narratives. Transitioning to Random Forest, we observed a notable uptick in accuracy (~83 %), precision (~81 %), recall (~82 %), and F1-score (~81 %). This model's capacity to capture nonlinear patterns across features is evident. SHAP analysis revealed that call duration, account balance, and certain linguistic tokens in the Bangla narrative were highly influential, demonstrating the model's ability to leverage both structured and textual data.

XGBoost further enhanced performance, achieving accuracy (~86 %), precision (~85 %), recall (~87 %), and F1-score (~86 %). Its superior gradient-boosting approach evidently sharpened minority-class detection, particularly important for early stroke or Alzheimer's risk identification. The marginal increase in AUC (~0.89) over Random Forest (~0.87) indicates better class separability. Together, these results affirm the added value of optimized ensemble learners in handling class imbalance and subtle feature interactions. The performance of the ANN (dense neural network with embeddings) was competitive, with an accuracy (~84 %), precision (~83 %), recall (~81 %), F1-score (~82 %), and AUC (~0.88). While slightly trailing XGBoost, the ANN's strength lies in its ability to integrate dense linguistic embeddings with structured clinical features in a unified representation. Fluctuations across cross-validation folds suggested sensitivity to hyperparameter selection, and early stopping and learning-rate scheduling helped maintain stability and generalization.

Most notably, the CNN-LSTM with attention model outperformed all others, achieving the highest accuracy (~90 %), precision (~89 %), recall (~91 %), F1-score (~90 %), and AUC (~0.93). The convolutional layers effectively captured local semantic patterns within the Bangla text, while the LSTM layers identified sequential trends in temporal features and clinical history. The attention mechanism added interpretability by spotlighting the most influential tokens and symptom phrases. This model's superior discriminative performance validates the hypothesis that fusing linguistic, imaging, and structured data enriches diagnostic insight. The AUC ranking across models reinforces this hierarchy: Logistic Regression (~0.75) at the bottom, Random Forest (~0.87) and XGBoost (~0.89) in the middle, ANN (~0.88) slightly behind XGBoost, and CNN-LSTM + attention at the top (~0.93). These findings confirm that hybrid architectures not only enhance accuracy but also offer superior sensitivity and specificity, crucial for medical diagnostics. The attention-based explanations further enhance trustworthiness by providing clinicians with transparent reasoning.

Future Work

Building on these promising results, several avenues for future exploration emerge. First, expanding dataset richness by incorporating MRI imaging features alongside Bangla text and structured variables would improve diagnostic accuracy in Alzheimer's and stroke prediction. Multi-source fusion remains untested; integrating imaging-derived biomarkers via CNNs or radiomics could yield stronger signals and improve early detection capabilities. Second, exploring transfer learning and pretrained language models fine-tuned on medical Bangla corpora may substantially enhance text representation quality. Adapting models such as multilingual BERT or domain-specific transformers could improve the semantic understanding

of clinical narratives and increase model robustness. Third, evaluating cross-lingual adaptability by extending the framework to other low-resource languages (e.g., Nepali, Urdu) would test its generalizability. Such multilingual, explainable architectures could serve diverse patient populations and improve inclusivity in AI-driven healthcare.

Fourth, refining explainability interfaces, developing interactive dashboards that integrate attention heatmaps, SHAP visualizations, and MRI highlights, would better meet clinicians' needs. User-centered evaluation through clinician interviews could surface actionable insights into usability, comprehension, and trust. Fifth, investigating privacy-preserving techniques, such as federated learning or encrypted inference, would allow deployment in data-sensitive healthcare environments. Ensuring model interpretability while preserving patient confidentiality is vital for real-world adoption. Finally, real-world prospective validation in clinical settings would be the ultimate test of model utility. Pilot deployment in hospitals that process Bangla text patient notes could provide objective feedback on operational performance, reliability, and integration challenges. To sum things up, the multimodal, hybrid explainable architecture demonstrated here offers strong diagnostic potential and interpretability. Future research should focus on data-scale enrichment, language adaptability, clinician-centered design, privacy frameworks, and real-world validation to fully realize its impact in diverse healthcare contexts.

7. Conclusion

This paper proposed a hybrid explainable AI framework that integrates Bangla textual narratives, structured clinical variables, and neural imaging representations to improve diagnostic prediction for stroke and Alzheimer's disease. By progressing from interpretable baselines through ensemble tree learners and dense neural networks to a CNN-LSTM architecture with attention, the work demonstrates that multimodal fusion yields meaningful gains in discriminative performance and clinical interpretability. The CNN-LSTM with attention achieved the strongest results across accuracy, precision, recall, F1, and AUC, indicating that local semantic patterns in Bangla text combined with temporal encoding and structured features provide the richest diagnostic signal. Tree ensembles such as Random Forest and XGBoost offered competitive performance while retaining clear feature attributions, and ANNs provided flexible joint representations of linguistic and tabular inputs. These outcomes support the core hypothesis that combining modalities and building explainability into the pipeline increases both predictive power and clinician trust.

At the same time, important limitations temper these findings and guide the next steps. The experimental results are derived from a prototype dataset and simulated pipelines rather than large-scale, clinically labeled Bangla corpora and multi-center MRI collections, so external validation is required before clinical deployment. Model sensitivity to hyperparameter choices and class imbalance management highlights the need for robust tuning and prospective testing. Future work should prioritize expansion to real multimodal datasets, fine-tuning language models on medical Bangla texts, inclusion of imaging biomarkers, privacy-preserving training paradigms, and clinician-centered evaluation of explanation utility. If these directions are pursued, the framework can evolve from a proof of concept into a deployable decision support tool that respects language diversity, preserves patient privacy, and delivers transparent, actionable insights for frontline healthcare providers.

References

[1] Abubakkar, M., Sharif, K. S., Ahmad, I., Tabila, D. M., Alsaud, F. A., & Debnath, S. (2025, June). Explainable Suicide Risk Prediction with DeepFusion: A Hybrid Intelligence Approach. In 2025 4th International Conference on Electronics Representation and Algorithm (ICERA) (pp. 455–460). IEEE.

[2] Ahad, M. A., et al. (2025). AI-Based Product Clustering for E-Commerce Platforms: Enhancing Navigation and User Personalization. International Journal of Environmental Sciences, 156–171.

[3] Ahmed, I., et al. (2025). Optimizing Solar Energy Production in the USA: Time-Series Analysis Using AI for Smart Energy Management. arXiv preprint arXiv:2506.23368.

[4] Aljunaid, S. K., et al. (2025). Secure and Transparent Banking: Explainable Federated Learning (XFL) Model for Financial Fraud Detection. (MDPI / conference preprint / ResearchGate entry), 2025.

[5] Cardaioli, M., Marangoni, L., Martini, G., Mazzolin, F., Pajola, L., Ferretto Parodi, A., Saitta, A., Vernillo, M. C., et al. (2025). FD4QC: Application of Classical and Quantum-Hybrid Machine Learning for Financial Fraud Detection. arXiv preprint arXiv:2507.19402.

[6] Deng, T., Bi, S., & Xiao, J. (2025). Transformer-Based Financial Fraud Detection with Cloud-Optimized Real-Time Streaming. arXiv preprint arXiv:2501.19267.

[7] El-Sappagh, S., Alonso, J. M., Islam, S. M. R., Sultan, A. M., & Kwak, K. S. (2021). A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. Scientific Reports, 11, Article 2660.

[8] Fariha, N., et al. (2025). Advanced fraud detection using machine learning models: Enhancing financial transaction security. arXiv preprint arXiv:2506.10842.

[9] Jahan, S. (2023). Explainable AI-based Alzheimer's prediction and management using multimodal data. PLOS ONE, 18(10), e0294253.

[10] Khan, M. A. U. H., et al. (2025). Secure Energy Transactions Using Blockchain Leveraging AI for Fraud Detection and Energy Market Stability. arXiv preprint arXiv:2506.19870.

[11] Khan, M. N. M., et al. (2025). Assessing the Impact of ESG Factors on Financial Performance Using an AI-Enabled Predictive Model. International Journal of Environmental Sciences, 1792–1811.

[12] Krishnamoorthy, M. V. (2024/2025). Data Obfuscation through Latent Space Projection (LSP) for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection. arXiv preprint arXiv:2410.17459.

[13] Liu, C., Tang, H., Yang, Z., Zhou, K., & Cha, S. (2025). Big Data-Driven Fraud Detection Using Machine Learning and Real-Time Stream Processing. arXiv preprint arXiv:2506.02008.

[14] Mahmud, T., Barua, K., Barua, A., Das, S., Basnin, N., Hossain, M. S., & Andersson, K. (2024). An explainable AI paradigm for Alzheimer's diagnosis using deep transfer learning. Diagnostics (Basel), 14(3), 345.

[15] Rahman, M., Debnath, S., Rana, M., Murad, S. A., Muzahid, A. J. M., Rashid, S. Z., & Gafur, A. (2023, February). Bangla Text Summarization Analysis Using Machine Learning: An Extractive Approach. In International Human Engineering Symposium (pp. 65–80). Singapore: Springer Nature Singapore.

[16] Uddin, M. M., Ahmad, I., Abubakkar, M., Debnath, S., & Alsaud, F. A. (2025, June). Interpretable Alzheimer's Disease Diagnosis Via CNNs and MRI: an Explainable AI Approach. In 2025 4th International Conference on Electronics Representation and Algorithm (ICERA) (pp. 641–646). IEEE.

[17] Vallarino, D. (2025). Detecting Financial Fraud with Hybrid Deep Learning: A Mix-of-Experts Approach to Sequential and Anomalous Patterns. arXiv preprint arXiv:2504.03750.

[18] Zamil, M. Z. H., Islam, M. R., Debnath, S., Mia, M. T., Rahman, M. A., & Biswas, A. K. (2025, April). Stroke Prediction on Healthcare Data Using SMOTE and Explainable Machine Learning. In 2025 13th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1–6). IEEE.