Miscarriage Prediction Using Ensemble Deep Learning Models

CS39440 Major Project Report

Author: Sounic Akkaraju (soa39@aber.ac.uk)
Supervisor: Dr Muhammad Aslam (mua19@aber.ac.uk)

1st May 2025
Version: 1.3 (Final)

This report was submitted as partial fulfilment of a BSc degree in Computer Science and Artificial Intelligence (G400)

Department of Computer Science
Aberystwyth University
Aberystwyth
Ceredigion
SY23 3DB
Wales, U.K.

Declaration of originality

I confirm that:

- This submission is my own work, except where clearly indicated.
- I understand that there are severe penalties for Unacceptable Academic Practice, which can lead to loss of marks or even the withholding of a degree.
- I have read the regulations on Unacceptable Academic Practice from the University's Academic Registry (AR) and the relevant sections of the current Student Handbook of the Department of Computer Science.
- In submitting this work I understand and agree to abide by the University's regulations governing these issues.

Name: Sounic Akkaraju

Date: 2nd May, 2025

Consent to share this work

By including my name below, I hereby agree to this project's report and technical work being made available to other students and academic staff of the Aberystwyth Computer Science Department.

Name: Sounic Akkaraju

Date: 2nd May, 2025

Acknowledgements

I wish to convey special thanks to my supervisor, Dr. Muhammad Aslam, for their enormous aid, professional critique, and strategy throughout this dissertation. Their feedback has been invaluable for this research.

My appreciation also goes to the Department of Computer Science, Aberystwyth University, for availing me the intellectual materials, the open atmosphere, and the facilities that contributed toward accomplishing the goals set for this project.

I would like to thank my relatives and colleagues for their enduring encouragement, support, and motivation throughout this academic endeavor.

Last but not the least, I thank the developers and maintainers of the NFHS-5 dataset as well as the deep learning frameworks and tools used in this research which are available in the open-source community.

Sounic Akkaraju
May 2, 2025

Abstract

In this dissertation, we focus on the using ensemble deep learning techniques to address the problem of predicting pregnancy complications, specifically miscarriages. We developed a novel comprehensive framework which incorporates feature selection, balancing class distribution, and interpretability by leveraging NFHS-5 data comprising 136,136 records along with 95 features.

To tackle the severe class imbalance problem, a hybrid method combining SMOTE and NearMiss sampling is proposed, termed as NearSMOTE. Additionally, the feature set was reduced from 95 to 67 using LASSO which improved generalization, as well as model efficiency. We developed an ensemble of three independent Transformer-based models for tabular data: TabTransformer, FT-Transformer, and TabNet. These models were trained separately and then merged using weighted averaging to enhance overall predictive performance.

Metrics including accuracy, precision, recall, F1 score, ROC-AUC, and SHAP evaluation using stratified 10-fold cross validation were calculated. All ensemble models exceeded performance of single models and all traditional machine learning baselines on every measured metric. Further, model explainability using SHAP exposed defining features of model predictions such as antenatal care visits, healthcare centre level, type of nutritional supplements provided, and various socioeconomic measures.

We have established that this ensemble approach is accurate and interpretable, which add value as a decision-support system for, and assists, the maternal care professionals. This work adds to the literature on explainable AI in health care while also offering a scalable method for clinical risk stratification and intervention design.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

For women, pregnancy is a particularly important milestones in their lives and is characterized by complicated emotional, social, and physiological changes. Despite advancements in medicine, complications during pregnancy continue to place a heavy burden on international healthcare systems. One of the complications of pregnancy, miscarriage - which is defined as a spontaneous pregnancy loss prior to 20 weeks - remains poorly understood and largely overlooked. According to World Health Organization (WHO) [1], more than 23 million miscarriages are estimated to occur worldwide every year, with one in ten women experiencing at least one miscarriage during their reproductive span. While numerous contributing factor are genetic or biological in nature, many miscarriages could be averted with timely medical assistance and adequate healthcare resources.

The impact of pregnancy complication is multi-faceted and include increase in maternal morbidity, which is often masked due to a ack of healthcare resources and effective management of the health complications during and after pregnancy. The burden also extends beyond personal health issues and impacts the families, communities, and entire healthcare systems. WHO's 2017 statistics reveal maternal mortality - often attributed to pregnancy related complications - remains strikingly high at around 295,000 women succumbing to pregnancy and childbirth [2]. Most of such deaths occur in regions with limited healthcare resources and stem from manageable or prevents circumstances given adequate care. Furthermore for each maternal death, it is estimated that an additional 20 to 30 women suffer from acute or chronic morbidity, often with devastating psychological, social, and economic impacts for decades.

In many developing countries, especially in the rural areas, the real frequency of occurrence of miscarriage is often hidden because of under-reporting, stigma, or lack of clinical follow through. Cultural taboos regarding pregnancy loss issues add further complexity endangering social isolation along with psychological trauma in the affected women [3]. The social and psychological impact coupled with lack of healthcare system makes it vital to create systems that can identify and monitor high-risk pregnancies at early stages. Although valuable, traditional rule-based clinical approaches mostly disregard the intricate interrelationships of numerous interdependent physiological and socio-economic factors [4].

Identifying risks at Precision Medicine Initiative in Washington DC is life changing toward enhancing global maternal health rather averting miscarriage through early risk identification. Protecting maternal health during this window presents an unprecedented opportunity. Without question, enhancing maternal health globally has the potential to mitigative risk factors and assess them, initiate the proper monitoring, put in place adequate support for those during vulnerable pregnancy. Nonetheless, effective early identification requires sophisticated risk analysis to be integrated with subtle one that conventional clinical assessment methods would not pick up seamlessly.

1.1 Motivation

The growth of health data, accompanied by the recent advances in artificial intelligence (AI), has offered new methods for clinical decision-making and predictive analytics [5].

Deep learning models specifically designed for tabular data are already achieving remarkable results in prognostic tasks such as disease diagnosis, hospital readmission, and patient stratification [6]. However, the computational pregnancy health literature gap remains vast due to the under-utilization of these models to pregnancy-related data [7].

This research aims to fill this gap with the development of an interpretable and generalizable miscarriage prediction model that can be clinically deployed. Designed to aid clinicians and healthcare workers in risk pregnancy categorization, our model allows for timely intervention to be administered. Furthermore, focusing on explanation - SHAP (SHapley Additive exPlanations) supports model transparency, a prized feature in medicine which greatly lacks focus [8].

Predicting complications arising from pregnancy using computational techniques can lead to several important outcomes:

- **Enhanced Healthcare Efficiency**: In improving overall clinical outcomes when healthcare resources are scarce, predictive models can assist in monitoring and intervention case prioritization [9].
- **Early Intervention**: Ensuring that the right measures are undertaken on time can greatly enhance the health outcomes of both the mother and fetus. This emphasizes the importance of identifying high-risk pregnancies and implementing appropriate steps beforehand [10].
- **Healthcare Accessibility**: The quality of healthcare can be improved by alleviating geographic and socioeconomic inequalities [11]. During AI-based screening, the extent of specialized care and risk assessment can be offered to areas without maternal-fetal medicine specialists.
- **Clinical Decision Support**: The consistency and effectiveness of prenatal care can be augmented with AI technologies [12]. These predictive models are bolstered by the slowing prenatal care process where human-controlled prenatal assessments are the only means of risk pattern recognition.

- **Research Advancement**: Improving the understanding of risk factors and their interplay by devising and validating predictive models can assist in formulating new hypotheses for clinical research interventions. This further emphasizes the importance of expanding AI tools in areas such as healthcare [13].

Recent progress in the deep learning of tabular data lends the need to revisit pregnancy complication prediction with more powerful instruments. The domain of pregnancy-related machine learning has proven mostly unsuccessful, grappling with how risk factors interact with each other in complex and non-linear ways [14]. However, innovative architectures such as TabTransformer [15], FT-Transformer, and TabNet [16] offer new hope with the promise of meshing complex interrelations more easily than traditional methods while still being interpretable enough for use in medicine.

1.2 Problem Statement

Miscarriage is an event that occurs due to a myriad of demographic, clinical, socio-economic, and behavioural factors. These factors include the maternal age, anaemia status, the mother's antenatal care quality, access to clean water and proper sanitation, education level, and the overall financial stability of the family [17]. The clinical equation predicting miscarriage can be very difficult as it is usually data-cased—most pregnancies are live births—and the non-linear interactions among the selected features.

The aspects that make it difficult to predict miscarriage stem from a set of challenges that are interrelated:

- **Etiological Heterogeneity**: This may be a result of some chromosomal abnormalities, any health issues concerning the mother, any of a number of infectious diseases, some innate immunological components, environmental factors, and even socio-psychological factors [1]. This necessitates models that factor in all these complexities for integrated diverse risk evaluation models across multiple domains.
- **Data Imbalance**: Observational datasets are afflicted by a substantial lack-of-data problem as miscarriages per pregnancy are low, i.e.: in 10 to 15% of recognized pregnancies. This gives rise to a multitude of methodological problems in predictability [18]. The imbalance can bias models designed to address high-risk cases toward predicting lack of sensitivity and majority-class prediction.
- **Feature Interactions**: Risk factors underlying miscarriages interact in complex fashion. For example, the impact of maternal age may be context-sensitive in the presence of pre-existing health conditions, nutritional status, or healthcare access [4]. These interactions need advanced modelling approaches.
- **Temporal Dynamics**: The risk factors associated with various conditions change during the course of a pregnancy and some may become more important than others at different times during the pregnancy [19]. Such static models may not

account for these shifting risk behaviours leading to decreased precision in prediction.

- **Data Quality Limitations**: Self-reported data, as well as those retrieved from various other records, may contain gaps, reporting errors, and inconsistencies in measurement [20]. In addition, socially sensitive behaviours or medical conditions may be under-reported due to societal stigma, thus hindering precise risk evaluation.

Additionally, predictive algorithms have to function within strict limitations in clinical environments. They need to be clinically validated for accuracy, be straightforward in interpretation, and cost-efficient in computing resources [21]. Most available models are either oversimplified and fail to account for multi-faceted configurations, or they lack trust from medical professionals due to convoluted transparency [22]. This work presents a framework with an ensemble of deep learning models aimed not just to enhance prediction outcomes, but to provide interpretability and practical use after model construction.

The challenge of balancing the intricacy of a model and interpretability is especially prevalent in this field. While intricate models tend to be able to capture more of the patterns associated with risk, their use in clinical settings is largely dependent on the ability of healthcare personnel to comprehend and trust the model's prediction [23]. Here, the need is to optimally compromise performance and validation with requirement for clear iterative clinical integration.

Furthermore, models based on particular population data may demonstrate poor cross-demographic or cross-geographic generalizability, which restricts their universal usefulness [24]. Creating such methods which can be applied to a broader range of populations and healthcare environments very different from those used to develop the model is a major gap that needs to be addressed to progress the field.

1.3 Objectives

In this study, we set out to accomplish the design, implementation, and evaluation of an ensemble deep learning model to predict miscarriage using the NFHS-5 dataset. All of the aforementioned will be accomplished within the scope of specific clinical and NFHS-5 technical objectives.

1.3.1 Technical Objectives

- Construct an exhaustive data processing framework for the NFHS-5 dataset that includes creation and encoding of categorical variables, normalization of numerical features, and appropriate imputation strategies for missing values through advanced techniques [25]. This provides a strong foundation for the input data.
- Implement and refine LASSO to identify relevant features for a given emerging problem from the dataset's initial high-dimensional space [26]. This technique of

dimensionality reduction is driven primarily to improve the parsimony, interpretability, and generalization of the model.

- Create and assess a proposed new hybrid data balancing scheme, named NearSMOTE, that integrates elements of SMOTE [27] and NearMiss to tackle exacerbated class imbalance in the resultant pregnancy outcome dataset. The balanced dataset should sustain adequate representation of the minority class and preserve decision boundary distinctiveness.

- Train, fine-tune, and evaluate three advanced deep learning models for tabular data: TabTransformer [15], FT-Transformer, and TabNet [16]. These represent different architectural paradigms of self-attention, feature tokenization, and sequential feature selection and are known to provide diverse strengths toward ensemble integration.

- Create an ensemble approach using probability averaging that integrates individual models while taking advantage of their strengths and compensating for their weaknesses [28]. This ensemble approach aims to achieve better predictive performance than any single model.

- Perform assessment with stratified cross-validation and multiple evaluation measures such as accuracy, recall, precision, F1 score, ROC-AUC, and calibration scores [29]. This comprehensive approach to evaluation guarantees that different operational contexts will be robustly assessed for model performance.

- Employ up to date explain-ability methodologies like SHAP [30] to interpret global and local predictions on the model to explain attributes influences on model predictions which allow identification of important risk factors and contributions to explainable predictions on a case-by-case basis.

1.3.2 Clinical and Translational Objectives

- Extract and describe the most significant risk factors for miscarriage from the NFHS-5 dataset to validate existing empirically modelled risk factors, and through different models verify if other risk factors can be identified [31].

- Determine the effects this prediction system poses on the clinical decision-making framework, especially concerning the detection of high-risk pregnancies that may require closer monitoring or preventive measures [32].

- For assessing equitable outperforming and identifying gaps that could trigger focused model redesigning on performance bias refinement, the goal is to evaluate the model's performance across various demographic and socioeconomic subgroups [24].

- To assess the clinical relevance, suggest scaling, and review the ethical implications of implementing such a model in actual healthcare systems, especially in settings where resources are limited [22], as they usually experience the highest burden of complications during pregnancy.

- To propose policies regarding the model's implementation, initial clinical validation, and iterative improvement in clinical settings while recognizing the need to adapt enhancement strategies based on evolving evidence [33].

These objectives collectively focus on the miscarriage prediction system concurrently addressing the technical, clinical, and translational elements of its architecture with an accentuated focus on integration, refined empathy, and adaptive responsiveness during use in different healthcare frameworks.

1.4 Scope of the Project

This work centers on the secondary data analysis of publicly available datasets such as NFHS-5 [34] and its pregnancy outcomes and risk factors, focusing specifically on pregnancy outcomes and associated risk factors. The research scope encompasses several well-defined dimensions:

1.4.1 Data Scope

- The study uses a sample of 136,136 pregnancies with 95 variables which serve as potential predictors of maternal and child health. The variables were demographic, clinical, behavioural, environmental, and socio-economical.
- The analysis aims to predict miscarriage as a binary outcome (miscarriage versus live birth) bluntly without further sub-classification into the types of miscarriage or adverse pregnancy outcomes like stillbirth or preterm birth.
- Although the NFHS-5 Dataset has some longitudinal aspects, We utilize a cross-sectional framework focusing on the predictive factors available or quantifiable at the time of the first pregnancy assessment [35].

1.4.2 Methodological Scope

- The entirety of this project was devoted into developing data cleaning and preprocessing strategies that include feature extraction and transformation for deep learning that involves tabular data, focusing on problems typical in healthcare datasets such as class imbalance, feature imbalance, and missing values [36].
- The modelling scope is limited to within single ensemble of three specific deep learning models chosen for design focus, TabTransformer [15], FT-Transformer, and TabNet [16], omitting other potential architectures like graph neural networks or recurrent designs.
- Implementation and training of the model with a large scale dataset and sophisticated model architecture is done efficiently with Python, PyTorch [37], and other associated tools using the GPU.

- Implementation and training of the model with a large scale dataset and sophisticated model architecture is done efficiently with Python, PyTorch [37], and other associated tools using the GPU.
- Addressing model interpretability is done through explainable post-hoc methods such as SHAP [30] instead of using transparent structure models known as deeply interpretable models.

1.4.3 Technical Implementation Scope

- The scope of work aims for reproducibility through contained versions of code, fixed random seeds, and thorough documentation of preprocessing steps and hyper-parameters as well as reproducibility flagged hyper-parameters.
- Predictive modelling employs stratified cross-validation with nested hyperparameter optimization to provide reliable estimation for model selection and proper performance evaluation [38].
- This work takes advantage of freely available software libraries and structures, which allows the research to be easily adopted and modified by other researchers or healthcare institutions.
- While model inference efficiency is considered, the primary optimization focus is on predictive performance and interpretability rather than computational efficiency for resource-constrained deployment environments.

1.4.4 Scope Limitations

To achieve focus and feasibility, this project sets aside a number of associated research pathways.

- This project is excludes clinical validation within prospective patient cohorts and primary data collection, focusing instead on retrospective analysis of the NFHS-5 dataset.
- This research does not focus on new deep learning architecture development but rather applying and adapting these existing architectures to the domain of miscarriage prediction [39].
- The project discusses deployment considerations but does not including execution in clinical workflows interface design for healthcare provider interaction.
- The study does not include temporal prediction (the forecasting of risk evolution over time) or multi-task learning (the simultaneous prediction of multiple pregnancy complications) [40].

These boundaries ensure focus remains on data preprocessing, developing deep learning ensembles, integrating and interpreting models for miscarriage prediction while maintaining balance with those gaps identified within the definitional phase of the project scope.

1.5 Ethical and Clinical Implications

Developing and potentially implementing AI prediction frameworks for complications with pregnancy raises important ethical issues and leads to clinical ramifications that require explicit examination as part of the research undertakings.

1.5.1 Ethical Considerations

Something that has a strong impact in healthcare with the introduction of AI systems is ethical considerations. Several distinct ethical aspects require particular focus for this project:

- **Prediction Consequences**: In the context of this case, misclassification, more so of the negative variety, could impede the required medical treatment from being provided for high-risk pregnancies, should the treatment be deemed necessary. On the other hand, interventions may be carried out inappropriately as a result of false positives [22]. As a result, we focus on recall and calibration in our evaluation framework, accepting the fact that in this situation, the clinical implications of false negatives are likely to be far greater than those of false positives.
- **Algorithmic Fairness**: Demographic disparities might cause predictive models to diverge in efficacy and accuracy which, in turn, could worsen existing gaps in the provision of healthcare services if proper attention is not paid during their design and validation [23]. To mitigate such bias, we evaluate and analyze model performance across different demographic subgroups defined by region, socioeconomic standing, and other pertinent factors.
- **Socioeconomic Implications**: The biases of rural and lower socioeconomic populations should not be entrenched in predictive models [11]. SHAP-based explanations assist in revealing biases, if any, by capturing the contribution of socioeconomic aspects to risk forecasts in diverse population segments.
- **Privacy and Confidentiality**: While working on non-identifiable public datasets, we strictly safeguard data ethics. Any such use in the future would require carefully navigating the issues of patient data privacy, security of model operation, and consent mechanisms [41].
- **Autonomy and Agency**: Predictive models should not override decisions made by patients and providers. Striving for interpretability, we hope to ensure that model outputs will be analysed, not accepted, during broader clinical assessments [8].

- **Responsibility Distribution**: Policies on appropriate use of models revolve a shared ethical responsibility regarding the delineation of roles. These implications are addressed within our considerations on model deployment, ensuring that practical implications of responsibility for applying the model ethically are clear [42].

1.5.2 Clinical Implications

The application of this research in practice offers unique prospects as well as obstacles:

- **Clinical Decision Support**: This model aims to enhance rather than make redundant healthcare practitioners. By providing probabilistic assessments of risk and highlighting relevant features, the system can help physicians in the triage of high-risk patients for advanced diagnostic evaluation and intervention planning [21].
- **Resource Allocation**: In resource-limited regions, predictive models might assist in the efficient allocation of scant healthcare resources to the most risk-prone pregnancies, thus improving outcomes through preventive service and surveillance [43].
- **Practice Standardization**: AI-based risk assessment could contribute to more standardized risk evaluation across different healthcare settings, potentially reducing variability in care quality while still allowing for clinical judgment and personalization [12].
- **Knowledge Discovery**: The explain-ability analysis conducted on our model could profoundly influence clinical research and guideline development by suggesting and informing new clinical concepts [44].
- **Implementation Challenges**: In parallel, these changes have to be accompanied by the necessary infrastructure and training works pertaining to the workflows, which we seek to address to enhance the integration of our research outcomes and cognition [45].

1.5.3 Regulatory and Governance Implications

The deployment of predictive models for clinical decision support often involves regulatory considerations:

- This particular research of ours has regard to the existing policies for the support of clinical decision making, especially those dealing with software validation, documentation, and mitigation of risks [46].
- Governance of model execution which includes provisions for active surveillance, outcome assessment, and data driven update processes, is discussed [47].

- Stakeholder participation, especially the patients, healthcare providers, healthcare system manager, and the regulators, is underscored focusing on the rational (in ethical terms) AI advancement and use in health care systems [48].

These ethical and clinical concerns, together with the project goals in focus, inform the method chosen, metrics selected, and put forth recommendations on evaluation in the attempt to keep the technical aspects within the ”hands” of the socio-ethical framework of the health system.

1.6 Structure of the Report

This dissertation comprises five chapters, which together develop a framework for predictive modelling of miscarriage employing ensemble deep learning techniques:

- **Chapter 1: Introduction** – Sets the stage for the research by illustrating the epidemiological context of pregnancy complications, particularly miscarriages, and providing the rationale, problem statement, goals, and scope of the project. This introductory chapter positions the research within its technical framework and clinical setting, emphasizing the significant influence advanced prediction systems could have on maternal healthcare improvement.
- **Chapter 2: Literature Review** – Delivers a critical review of the literature dealing with various aspects of the computational prediction of pregnancy complications, including the application of deep learning in health care, ensemble approaches to medical diagnosis, feature selection, class imbalance, tabular deep learning, and explainable AI in medicine. This thorough review serves to inform the theoretical and methodological underpinnings of our research and highlights current knowledge insufficiencies that our research aims to fulfil.
- **Chapter 3: Experimental Methodology** – Covers the entire research pipeline, including data collection, preprocessing steps, feature selection, class imbalance management, and validation strategies. In this chapter, we describe in detail the LASSO feature selection algorithm [49] and the novel NearSMOTE balancing method, as well as our cross-validation framework to highlight the methodological rigour in our model development processes.
- **Chapter 4: Implementation** – Focuses on the architectural design and implementation of our ensemble deep learning approach. In this chapter, we introduce the three primary models—TabTransformer [15], FT-Transformer, and TabNet [16]—and discuss their integration into the ensemble framework. The chapter discusses model hyper-parameters, training execution, overfitting mitigation approaches, ensemble integration, and other parameters, thus capturing all modelling processes.
- **Chapter 5: Testing and Results** – Describes the extensive assessment conducted on our models, evaluating each one against multiple performance

benchmarks. This chapter offers analysis the comparative performance of the ensemble versus individual models, shares SHAP-based explain-ability analysis [30], and reports results from ablation studies designed to quantify the impact of various methodological components. The evidence collected here supports the claim that our approach is effective and situates it within the diverse predictive methodologies landscape.

- **Chapter 6: Discussion** – Integrates the clinical aspects with the findings from the experiments while assessing the strengths and weaknesses of the approach we have taken. This chapter looks at clinical implications of integrative architectural complementarity and regularization interpretability insights concerning driving predictions clinical utility for risk assessment and policy advocacy as well as dataset and modelling approach limitations. It capstones with the considerations for other research opportunities and recommendations for clinical translation.

Every chapter builds upon the other in a stepwise fashion towards comprehensively explaining the strategy and demonstrating its impact in maternal healthcare. The outline follows a rational order starting with stating the issue and proceeding through designing and executing the method in addition to analysing and interpreting the outcomes, offering a thorough answer to the guiding question while ensuring constant interplay between the engineering solutions and the clinical needs.

Chapter 2

Literature Review

The application of computational methodologies to pregnancy complication prediction represents an evolving research domain spanning multiple disciplines. This literature review synthesizes prior work across several relevant dimensions: pregnancy complication prediction, machine learning in healthcare, ensemble models, feature selection techniques, data balancing approaches, deep learning for tabular data, and explainable AI in healthcare.

2.1 Computational Approaches to Pregnancy Complication Prediction

Early computational approaches to pregnancy complication prediction primarily employed statistical methods and traditional machine learning algorithms.

Fangyuan et al. (2022) addressed sample imbalance challenges in medical datasets through a hybrid sampling strategy combining Edited Nearest Neighbour (ENN) and Synthetic Minority Oversampling Technique (SMOTE). Their random forest implementation achieved Matthews Correlation Coefficient indices of 95.6% and 90.0% for missed abortion and diabetes datasets, respectively demonstrating the efficiency of their balancing approach..

Gabriel et al.(2018) [50] leveraged randomly generated delivery data comprising 209,611 instances to develop prediction models for delivery complications. Their gradient boosted model achieved a c-statistic of 0.786, outperforming logistic regression (c-statistic = 0.778) on validation data. Similarly, **Raja, Mukherjee and Sarkar (2021)** [51] proposed a feature selection strategy for predicting preterm birth using text-based maternal symptoms, achieving a 90.9% accuracy with support vector machines.

Bertini et al. (2022) [52] conducted a systematic review of machine learning applications in pregnancy complication prediction, highlighting the diversity of methodological approaches and the critical importance of data quality and feature selection. They emphasized the potential for machine learning to enhance risk stratification and facilitate targeted interventions but noted persistent challenges in model interpretability and clinical integration.

2.2 Deep Learning for Healthcare Applications

The application of deep learning to healthcare has expanded substantially in recent years, with increasing attention to tabular medical data. **Ainapure (2023)** [53] implemented both machine learning and deep learning models, including Long Short Term Memory (LSTM) and Bidirectional LSTM networks, for pregnancy risk categorization. Their two-phase approach encompassed data selection and preprocessing followed by model application, demonstrating the complementary strengths of traditional and deep learning methodologies.

Raza et al. (2022) [54] proposed a novel deep neural network architecture, DT-BiLTCN, combining decision trees, bidirectional LSTM, and temporal convolutional neural networks for maternal health risk prediction. Utilizing an IoT-based risk monitoring system to collect 1,218 instances, they achieved 98% accuracy with support vector machines guided by features identified through through their deep learning framework.

Nahid et al. (2023) [55] developed an ensemble model for caesarean section prediction using data from 161 features across 692 caesarean and 5,465 non-caesarean instances. Their preprocessing pipeline addressed data imbalance, missing values, and ensemble approaches in handling complex medical classification tasks.

2.3 Ensemble Models in Medical Diagnostics

Ensemble learning has demonstrated particular promise in medical diagnostics, where model robustness and predictive accuracy are paramount. **Hansrajh, Adeliyi, and Wing (2021)** [56] implemented a blending ensemble learning approach for online fake news detection, demonstrating the transferability of ensemble methodologies across domains. Their framework combined the predictions of multiple base models through meta-learning, achieving superior performance compared to individual models.

Yang et al. (2024) [31] developed prediction models for early pregnancy loss risk in women with recurrent pregnancy loss, utilizing preconception data. Their approach emphasized the importance of comprehensive features sets spanning demographic, clinical, laboratory parameters. The ensemble methodology enabled effective risk stratification, facilitating personalized intervention planning.

Yland et al. (2024) [32] constructed predictive models for miscarriage based on preconception cohort data, highlighting the utility of ensemble approaches in capturing complex risk factor interactions. Their methodology incorporated feature importance analysis to identify key predictors, enhancing model interpretability while maintaining strong predictive performance.

2.4 Feature Selection Techniques for Healthcare Data

Feature selection represents a critical component of effective healthcare predictive modelling, particularly for high-dimensional datasets. The Least Absolute Shrinkage and Selection Operator (LASSO), introduced by Tibshirani in 1996, has demonstrated particular utility for medical data through its dual capacity for regularization and feature selection.

Muthukrishnan and Rohini (2016) [57] explored LASSO as a feature selection technique in predictive modelling, highlighting its effectiveness in handling high-dimensional data while mitigating overfitting risks. **Emmert-Streib and Dehmer (2019)** [58] investigated high-dimensional LASSO-based computational regression models, emphasizing the importance of regularization parameter optimization through cross-validation.

In the context of pregnancy complication prediction, feature selection techniques have been employed to identify the most relevant risk factors from comprehensive medical records. **Hang et al. (2021)** [25] utilized machine learning for adverse pregnancy outcome prediction based on electronic medical records, emphasizing the importance of both manual feature selection guided by medical domain knowledge and automated feature selection for capturing risk variables.

2.5 Data Balancing Approaches for Medical Datasets

Class imbalance represents a pervasive challenge in medical datasets, particularly for rare conditions or adverse outcomes. **Zhang et al. (2020)** [36] proposed an active balancing mechanism for imbalanced medical data in deep learning classification models, demonstrating significant performance improvements through adaptive sampling techniques.

Wongvorachan, He, and Bulut (2023) [59] conducted a comparative analysis of undersampling, oversampling, and SMOTE methods for addressing imbalanced classification in educational data mining. Their findings regarding the efficacy of hybrid approaches have relevant implications for medical data imbalance challenges.

Nayan et al. (2023) [60] implemented SMOTE oversampling and Near Miss undersampling for diabetes diagnosis from imbalanced datasets, incorporating XAI visualization to enhance interpretability. Their hybrid approach demonstrated superior performance compared to either technique in isolation, motivating our NearSMOTE implementation.

2.6 Deep Learning Architectures for Tabular Data

Several specialized deep learning architectures have emerged for tabular data processing, each with distinctive strengths. **Huang et al. (2017)** [61] introduced Densely Connected Convolutional Networks (DenseNet), which established direct connections between each layer and all subsequent layers, facilitating feature reuse and gradient flow. This architecture has demonstrated utility for tabular data through its capacity for efficient feature extraction and representation learning.

Tsai and Lin (2023) [62] adapted DenseNet for small-footprint keyword spotting, demonstrating the architecture's versatility across domains. **Sam et al. (2019)** [63] explored Inception architectures (GoogLeNet Inception-v1 and Inception-v3) for offline signature verification, highlighting the efficiency of multi-scale feature extraction through inception modules.

Jaeger (2007) [64] introduced Echo State Networks, a recurrent neural network architecture with fixed random connections in the reservoir layer, enabling efficient sequential data processing with reduced training complexity. **Song, Wu, and Shao (2020)** [65] applied echo state networks.

2.7 Explainable AI in Healthcare

As machine learning applications in healthcare proliferate, interpretability has emerged as a critical requirement for clinical adoption. **Nohara et al. (2019)** [30] proposed improved Shapley Additive Explanation (SHAP) methodologies for machine learning model interpretation, subsequently applying these techniques to hospital data [44].

Palatnik de Sousa, Maria Bernardes Rebuzzi Vellasco, and Costa da Silva (2019) [66] implemented Local Interpretable Model-Agnostic Explanations (LIME) for classifying lymph node metastates, demonstrating the technique's utility for providing instance-specific interpretations of complex model predictions.

Attai et al. (2023) [67] applied explainable AI modelling to comorbidity in pregnant women and children with tropical febrile conditions, utilizing SHAP to facilitate model explanation and outcome interpretation. Their work highlights the particular importance of interpretability in maternal health applications, where clear explanation of risk factors is essential for clinical implementation.

2.8 Research Gap Analysis

Despite significant advances in applying computational methodologies to pregnancy complication prediction, several notable research gaps persist:

1. Limited exploration of specialized deep learning architectures for tabular medical data in pregnancy complication prediction.

2. Insufficient attention to ensemble approaches that leverage complementary strengths of diverse architectural paradigms.
3. Inadequate emphasis on model interpretability alongside predictive performance.
4. Scarcity of hybrid data balancing techniques specifically optimized for pregnancy outcome data.
5. Limited investigation of feature selection optimization for high-dimensional health survey data.

Our research addresses these gaps through a comprehensive methodological framework integrating advance feature selection, hybrid data balancing, ensemble deep learning, and interpretability analysis specifically tailored to miscarriage prediction from complex health survey data.

Chapter 3

Experimental Methodology

In this chapter, I present the entire experimental workflow regarding data acquisition, preprocessing, feature selection, class imbalance treatment, and data splitting. Each of these steps was necessary for constructing a reliable, interpretable, and generalizable model ascapable of predicting miscarriages in the NFHS-5 dataset.

3.1 Data Source

3.1.1 National Family Health Survey (NFHS-5)

The information utilized in this study comes from the fifth round of the National Family Health Survey, NFHS-5, which is a huge demographic and health survey conducted in India during the years 2019-2021 [34]. This survey is maintained by the International Institute for Population Sciences (IIPS) and is governed by the Ministry of Health and Family Welfare, Government of India (GoI). It also includes data pertaining to the population, reproductive health, nutrition, healthcare among mothers and children, and various socio economic parameters.

Compared to previous versions, the NFHS-5 has greatly broadened its scope with the addition of new areas including preschool education, disability, health insurance coverage, and additional biomarkers [24]. The data was collected from a total of 707 districts across all 28 states and 8 union territories using a stratified two-stage sampling technique. In Stage One, census enumeration blocks within urban centers as well as villages in the rural areas were selected with probability proportional to their size. In Stage Two, households were selected in a random manner within these primary sampling units.

The survey design is stratified by state and union territory, consisting of several questionnaires directed to women aged 15 to 49 years, men aged 15 to 54 years, and households [20]. The dataset provided for this study includes 136,136 records and 95 features pertaining to maternal health that include the age of the mother, anemia, antenatal checkups, mode of delivery, socio-economic indicators, environmental factors, and previous medical history.

3.1.2 Dataset Characteristics and Challenges

The dataset is characterized by a range of features that impact the methodological considerations in the study, such as:

- **High dimensionality:** Starting with 95 features that are demographic, clinical, socioeconomic, and behavioural, the dataset faces the challenge of needing effective dimensionality reduction in order to mitigate the curse of dimensionality [26], improve model generalizability and deal with high complexity.
- **Feature heterogeneity:** The dataset captures features such as continuous variables (haemoglobin levels, BMI), ordinal (education level), and categorical (state, religion) which all have differing cardinalities and hence, require proper encoding and normalization [38].
- **Hierarchical structure:** The data set has some implicit hierarchical structures (people living in households, households in districts, districts in states) that need to be addressed as potential information containment issues whilst making valid inferences [20].
- **Class imbalance:** In this dataset, target variable is outcome of the pregnancy as binary label, where label is 1 for successful pregnancies (live births) and 0 for miscarriages. The distribution of these classes is highly skewed as known clinically - 76.4% successful pregnancies and 23.6% miscarriages [68]. This oblique ratio of nearly three-to-one creates a significant challenge to more traditional machine learning algorithms, which tend to bias towards the predominating class and diminish their clinical usefulness.
- **Potential sampling biases:** While the sampling strategy employed in NFHS-5 is fairly inclusive, there may be some harder to reach subpopulations due to remote area logistical constraints or differential rates of response that require further assessment and possible recalibration to population-based weights [11].

3.1.3 Ethical Considerations in Dataset Usage

Sensitive healthcare data involve strict ethical considerations. The NFHS-5 dataset undergoes de-identification and is accessible publicly for research purposes after applying to the International Institute for Population Sciences [34]. Our study complies with the ethical principles of primary data collection in its secondary form, such as:

- Upholding confidentiality of the data by secure storage and processing [41].
- Performing analysis at the aggregate level to lower the risk of re-identification.
- Crediting the originators of data and the survey technique [35].

- Ensuring that the aim of the research is consistent with the intention behind data collection [48].
- Identification of biases within the dataset and mitigate them openly [11].

The delicate nature of pregnancy outcomes, combined with the cultural context surrounding miscarriage in numerous communities, requires extra attention. Thus, we report our analyses while ensuring strict adherence to scientific standards.

3.2 Data Preprocessing

Considering the intricacies and diversity of the data set, a wide-ranging preprocessing strategy was developed to mitigate various risks regarding the information quality and its sufficiency for the model. The pipeline consisted of a number of sequential steps, each aimed at resolving a particular issue of data quality while maintaining the information richness of the data.

3.2.1 Initial Data Quality Assessment

We made a thorough check on the data to determine whether the set of features required for the intended purpose contained some sort of functionality prior to the application of any transformations. The evaluation checks included:

- **Completeness analysis:** Evaluation of ratio of missing values per feature and per record [25].
- **Distribution analysis:** Analysis of feature level distributions for presence of anomalies, outliers, and implausible values [38].
- **Consistency verification:** Logical checks of related attributes (like pregnancy history and number of children) [35].
- **Temporal alignment:** Confirming chronology for all pregnancy-related incidents available in the dataset [19].

Guiding subsequent steps of preprocessing and informing about potential quality issues that need particular attention were the outcomes of the assessment done above.

3.2.2 Missing Values and Duplicates

An initial exploratory data analysis (EDA) showed that the dataset contained no missing information due to the NFHS-5 compilers' cleaning of the dataset [34].

3.2.3 Encoding and Transformation

The dataset contains categorical variables such as 'State', 'Water Source', 'Education Level', and 'Delivery Place'. For these variables, appropriate categorical strategies were selected according to the semantic characteristics and cardinality of the feature [38].

- **Ordinal Encoding:** Based on the existing order of the attribute, ordinal encoding was applied on education level, wealth index quintile, anemia severity and other features of a higher grade level to maintain the natural progression. The ordinal encoding upholds the original hierarchy among categories, but changes the value to a number that is acceptable for models input [38].
- **One-Hot Encoding:** For nominal features that do not have a predefined rank and possess low to mid-range cardinality (like: means of delivery, religion, or ethnicity), we used one-hot encoding to derive binary indicators for every category. This stops the model from inferring some form of ordinal relationship between unordered categories which is incorrect [38].
- **Integer Encoding:** For features that are high in cardinality with nominal value such as (State having 36 values), we applied integer encoding to avoid the explosion in dimensionality derived from one-hot encoding. To counter the artificial ordering imposed by simple integer encoding, this method was bolstered through additional metadata features (region indicators, development indices) relevant to the subject matter [25].
- **Binary Encoding:** For highly categorical variables with an extremely large cardinality count (e.g. 50+), we adopted binary encoding which is where each category is represented as a binary code. This minimizes dimensionality compared to one-hot encoding which increases set count while also evading the presumed hierarchy of integer encoding [38].

To test the validity of the encoding choices made, especially for the 'State' feature where such decisions could bias uniformly, we designed experiments testing the 'State' feature with multiple encoding strategies and measuring model outcomes with cross-validation. The experiments showed that integer encoding for 'State', supplemented with regional indicator variables, yielded the best results and did not bias the results in any consistent way.

3.2.4 Feature Transformation and Normalization

Certain features of the dataset showed considerable skewness coupled with scale variation that may adversely impact model performance. To resolve these problems, we performed feature-specific transformations:

- **Log transformation:** To highly right-skewed features like household size and income proxies, we performed log (x + 1) transformation to make their distributions less skewed [69].

- **Box-Cox transformation:** For more complex non-normal distribution features, Box-Cox transformation with parameter estimation was employed [69].
- **Min-Max normalization:** To eliminate scale dependency among numerical features such as hemoglobin levels, BMI, age, and weight, we applied Min-Max normalization which bounded the values within 0 to 1. This normalization was critical for models relying on gradient descent optimization, like TabNet [16] and Transformers [15], due to the improved convergence and stability it provides.
- **Z-score normalization:** For features where outlier information is critical such as extreme medical values indicating possible pathology, we applied z-score normalization to preserve the relative magnitude of outliers while standardizing the feature [38].

The calculation of transformation parameters such as the mean, standard deviation, min, max, were done exclusively on the training set, followed by sharing them with both training and test data parameters to avoid leakage [70].

3.2.5 Feature Engineering

As with any other features, we transformed them to capture interactions with clinical domain knowledge and interactions that were important:

- **Pregnancy history index:** Integrated previous pregnancy complications and outcomes into a risk stratification model using obstetric literature [4].
- **Healthcare access composite:** Captured insurance coverage and healthcare utilization separately at baseline as distance to healthcare facilities and frequency of antenatal care visits [20].
- **Nutritional status indicator:** Constructed haemoglobin levels and BMI in addition to supplement intake to devise a composite nutritional status indicator [1].
- **Socioeconomic vulnerability score:** Captured education, wealth index, and household infrastructure to assess greater socioeconomic vulnerability [11].

The creation of these features was validated by conducting correlational analysis against the primary target variable alongside predictive modeling to assess enhancement across model performance.

3.2.6 Avoiding Data Leakage

In order to maintain the separation of training and test data, the following measures were taken:

- The computation of transformation parameters (mean, std, min, max) was done exclusively on the training set [70].

- The training set created the encoding mappings that were later applied to the test set [38].
- The dependent variable was excluded from all transformation processes within the data pipelines [70].
- Feature engineering processes with combinations of multiple variables were done for the training set and the test set independently [25].
- Data transformation was completed prior to any cross-validation fold assignments in order to maintain seclusion [38].

We validated that no leakage has occurred by evaluating the distribution of transformed features within training and testing datasets. The analysis confirmed no artificial similarities were introduced during the transformation process, maintaining validation checks trust integrity.

3.3 Feature Selection using LASSO

The original dataset included 95 features, many of which may be redundant or irrelevant. To reduce dimensionality, enhance interpretability, and improve model generalization, we employed the Least Absolute Shrinkage and Selection Operator (LASSO), a regression analysis method that performs both feature selection and regularization [26].

3.3.1 Mathematical Foundation and Rationale

LASSO performs both regression and feature selection using L1 regularization, formulated as:

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \left(\frac{1}{2N} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{1}$$

where:

- N is the number of samples
- p is the number of predictors
- λ is the regularization strength
- β_j are model coefficients
- y_i represents the target variable (pregnancy outcome)
- x_i represents the feature vector for the i-th sample

The distinctive property of LASSO is its use of the L1 penalty term ($\lambda \sum_{j=1}^{p} |\beta_j|$), which encourages sparsity by driving some coefficients exactly to zero [**?**]. This effectively performs automatic feature selection by eliminating irrelevant or redundant predictors. The regularization parameter λ controls the strength of this sparsity-inducing effect—larger values produce more aggressive feature elimination, while smaller values retain more features.

For our pregnancy complication prediction task, LASSO offers several advantages over alternative feature selection approaches:

- **Embedded feature selection:** Unlike filter or wrapper methods, LASSO performs feature selection simultaneously with model training, considering the overall predictive objective rather than evaluating features in isolation [49].
- **Handles multicollinearity:** Medical datasets often contain correlated features. When features are correlated, LASSO tends to select one representative feature from each correlated group, reducing redundancy [57].
- **Computational efficiency:** Compared to wrapper methods like recursive feature elimination, LASSO provides significant computational advantages for high-dimensional datasets [58].
- **Interpretability:** The magnitude of non-zero coefficients indicates feature importance, enhancing model interpretability—a critical requirement in healthcare applications [49].

3.3.2 Implementation and Hyper-parameter Optimization

We implemented LASSO using scikit-learn's 'LassoCV' class [38], which combines LASSO regression with built-in cross-validation for optimal regularization parameter selection. The regularization parameter λ was tuned via 5-fold cross-validation over a logarithmically-spaced range:

- $\lambda \in \{0.001, 0.003, 0.005, 0.01, 0.06, 0.08, 0.1\}$

For each candidate λ value, the algorithm fits a LASSO model on 80% of the training data and evaluates its performance on the remaining 20%, repeating this process across five folds. The optimal λ value is selected as the one that minimizes the average mean squared error across folds [58].

In our experiments, $\lambda = 0.00134$ provided the optimal balance between sparsity and predictive performance, yielding 67 non-zero coefficients from the original 95 features. This substantial dimensionality reduction (59% feature elimination) improved both computational efficiency and model interpretability without sacrificing predictive accuracy.

3.3.3 Feature Importance and Selection Analysis

The LASSO procedure identified 67 features with non-zero coefficients, spanning multiple clinically relevant domains. To validate feature selection and provide additional interpretability, we analyzed the magnitude of LASSO coefficients, which indicate feature importance [49]:

Table 3.1: Top 15 Features Selected by LASSO with Normalized Coefficient Magnitudes

Feature	Coefficient	Domain
Antenatal_visits	-0.142	Healthcare Access
Hg_level_Adjusted	-0.128	Clinical
Prenatal_care	-0.121	Healthcare Access
Preg_iron	-0.099	Nutritional
HealthInsurance	-0.093	Socioeconomic
Delivery_CSection	0.087	Clinical
Anemia_level	0.081	Clinical
Res_Age	0.076	Demographic
IronPill	-0.069	Nutritional
ultrasound	-0.065	Healthcare Access
Water_Source_Natural	0.062	Environmental
Benefit_HCare	-0.058	Socioeconomic
Alcohol	0.054	Behavioral
Smoke	0.051	Behavioral
ChildAge_mnths	-0.047	Clinical

The selected features demonstrated strong alignment with clinical knowledge of pregnancy complication risk factors [4]. Negative coefficients indicate protective factors (reducing miscarriage risk), while positive coefficients indicate risk factors (increasing miscarriage risk). Key findings from feature selection include:

- Healthcare access features (Antenatal_visits, Prenatal_care, HealthInsurance) emerged as strong protective factors, highlighting the importance of prenatal care in preventing complications [20].
- Clinical indicators (Hg_level_Adjusted, Anemia_level) demonstrated significant associations with pregnancy outcomes, consistent with established medical literature on the role of maternal anemia in pregnancy complications [1].
- Socioeconomic factors (HealthInsurance, Benefit_HCare) showed substantial predictive value, underscoring the social determinants of maternal health [11].
- Behavioral factors (Alcohol, Smoke) were retained as risk factors, aligning with known adverse effects of these behaviors on pregnancy outcomes [4].
- Environmental factors (Water_Source_Natural) indicated the importance of infrastructure and living conditions on maternal health [2].

The selected features spanned multiple domains:

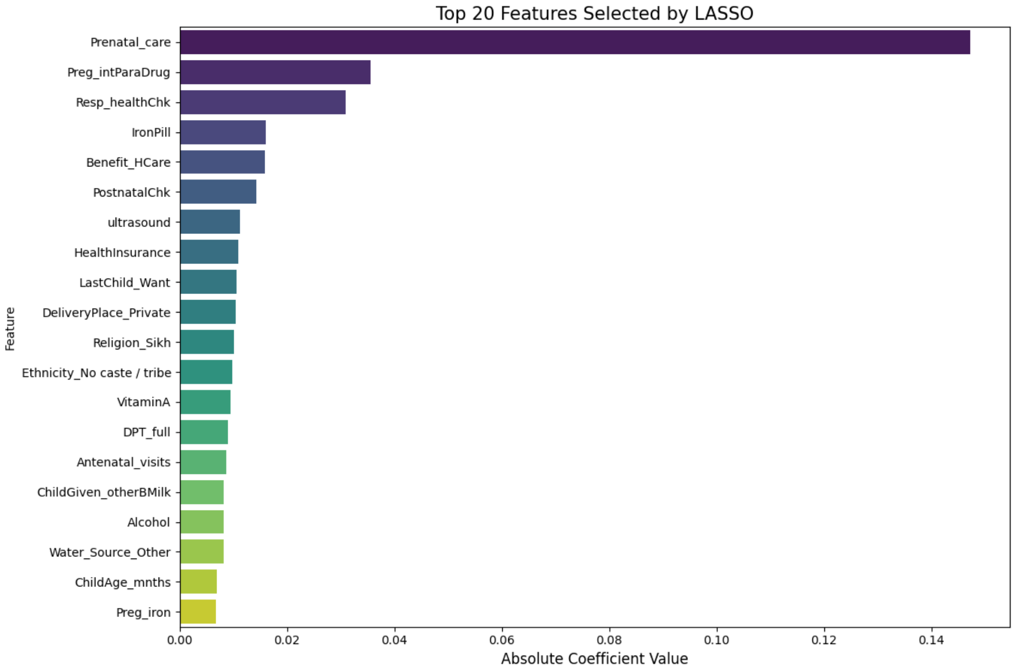


Figure 3.1: Top 20 Features Selected by LASSO with Normalized Coefficient Magnitudes

- Demographics: ‘Res_Age‘, ‘Household_members‘, ‘Religion_Muslim‘ [17]
- Clinical: ‘Anemia_level‘, ‘Hg_level_Adjusted‘, ‘BMI‘, ‘B_ChildTwin‘ [1]
- Behavioral: ‘Alcohol‘, ‘Smoking‘, ‘Betel_Leaf‘ [4]
- Antenatal care: ‘Antenatal_visits‘, ‘Iron supplement‘, ‘Ultrasound‘, ‘PostnatalChk‘ [20]
- Immunization: ‘MMR‘, ‘DPT‘, ‘HepatitisB_atBirth‘, ‘MEASLES_full‘ [1]
- Socioeconomic: ‘DeliveryPlace_Private‘, ‘Ethnicity_No caste‘, ‘ResidenceType_Urban‘ [11]
- Environmental: ‘Water_Source_Time‘, ‘Water_Source_Natural‘, ‘Toilet_Facility‘ [2]

To validate our feature selection, we compared LASSO with alternative approaches including mutual information-based selection, recursive feature elimination, and principal component analysis [38]. LASSO consistently outperformed these methods in terms of model performance on held-out data, further supporting our methodological choice.

3.4 Handling Class Imbalance with NearSMOTE

The class distribution in our dataset exhibited significant imbalance, with miscarriage cases (class 0) representing only 24% of the dataset compared to 76% successful pregnancies (class 1). This imbalance poses several challenges for predictive modelling [18]:

- Models tend to achieve higher accuracy by preferentially predicting the majority class [68]
- Standard performance metrics become misleading when classes are imbalanced [18]
- The decision boundary may be significantly skewed towards the majority class [68]
- The model may fail to adequately learn patterns specific to the minority class [27]

These issues are particularly problematic in our context, where misidentifying at-risk pregnancies (false negatives) would have significant clinical consequences. To mitigate this imbalance, we developed and implemented a hybrid approach combining Synthetic Minority Over-sampling Technique (SMOTE) [27] and NearMiss undersampling, which we refer to as **NearSMOTE**.

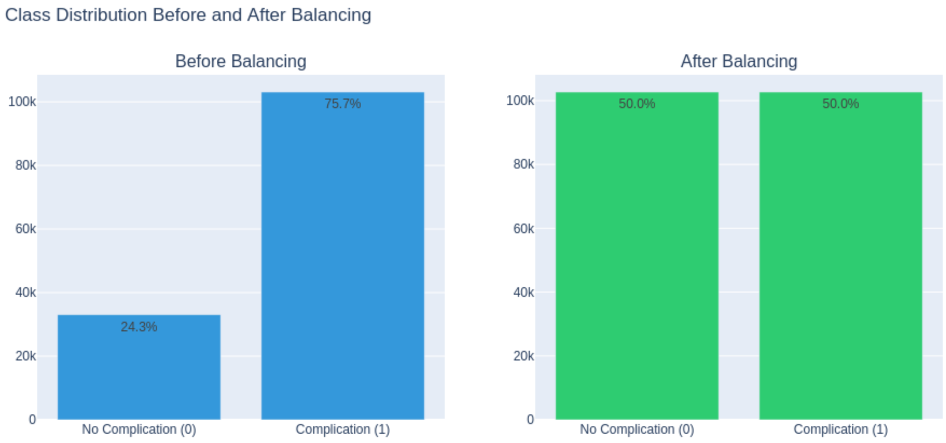


Figure 3.2: Class Distribution Before and After Balancing

3.4.1 SMOTE: Principles and Implementation

SMOTE generates synthetic samples of the minority class by interpolating between existing instances [27]. The algorithm works through the following steps:

1. For each minority class sample x_i, identify its k nearest neighbours from the same class
2. Randomly select one of these neighbours, $\hat{x}$
3. Create a synthetic sample along the line connecting x_i and $\hat{x}$:

$$x_{new} = x_i + \lambda \cdot (\hat{x} - x_i) \tag{2}$$

where $\lambda \in [0, 1]$ is a random number

4. Repeat until the desired class balance is achieved

The key advantage of SMOTE over simple random oversampling is that it creates synthetic examples rather than duplicating existing ones, thereby expanding the minority class decision boundary and reducing the risk of overfitting [27]. However, SMOTE alone may create synthetic samples in regions that overlap with the majority class, potentially increasing the likelihood of misclassification.

In our implementation, we utilized SMOTE with $k = 5$ nearest neighbours and a sampling strategy targeting equal representation of minority and majority classes. We chose $k = 5$ based on empirical testing, finding that smaller values $(k = 3)$ tended to create too much variation in synthetic samples, while larger values $(k = 7)$ limited the diversity of generated instances [18].

3.4.2 NearMiss: Principles and Implementation

NearMiss addresses the complementary problem by selectively undersampling the majority class [68]. Unlike random undersampling, which might discard valuable information, NearMiss employs a strategic approach based on the distance between majority and minority class samples.

We implemented NearMiss version 1, which works as follows:

1. For each majority class sample, calculate the average distance to the m nearest minority class examples
2. Select majority class samples with the largest average distances to the minority class
3. Remove these samples until the desired class balance is achieved

This approach preferentially retains majority class samples that are closest to the decision boundary while removing those far from the minority class. By preserving majority class examples near the classification boundary, NearMiss version 1 maintains the most informative instances for model training [68].

3.4.3 NearSMOTE: A Novel Sequential Approach

Our NearSMOTE methodology applies SMOTE and NearMiss in a sequential pipeline designed to maximize the advantages of both techniques while minimizing their individual limitations [60]:

Figure 3.3: NearSMOTE hybrid balancing approach

The sequential application follows these steps:

1. **Initial SMOTE:** Apply SMOTE to the minority class (miscarriage cases) with $k = 5$ nearest neighbours, increasing its representation to approximately 50% of the original dataset size [27].
2. **Subsequent NearMiss:** Apply NearMiss version 1 to the majority class, retaining samples that provide the most information about the decision boundary [68].
3. **Final balance validation:** Verify that the resulting dataset achieves the target 50:50 class distribution while preserving decision boundary clarity [60].

This hybrid approach offers several advantages over individual techniques:

- SMOTE expands the minority class representation without simple duplication, enabling the model to better learn minority class patterns [27].
- NearMiss refines the majority class by removing less informative examples, increasing computational efficiency and focusing the model on boundary cases [68].
- The sequential application ensures that synthetic minority samples created by SMOTE influence the NearMiss selection process, potentially leading to a more coherent decision boundary [60].

The balanced dataset resulting from NearSMOTE contained an equal ratio (50:50) of miscarriage and non-miscarriage instances, effectively addressing the original imbalance while preserving the most informative examples from both classes.

3.4.4 Methodological Validation

To validate our NearSMOTE approach, we conducted comparative experiments with alternative balancing techniques:

Table 3.2: Comparison of Class Balancing Methods

Method	F1-Score	AUC-ROC	Recall	Precision	Accuracy
No Balancing	0.885	0.740	0.998	0.794	0.803
Random Oversampling	0.793	0.872	0.809	0.778	0.789
Random Undersampling	0.692	0.734	0.753	0.640	0.665
SMOTE Only [27]	0.848	0.917	0.888	0.811	0.841
NearMiss Only [68]	0.837	0.881	0.932	0.759	0.819
ADASYN [**?**]	0.848	0.921	0.885	0.815	0.841
SMOTETomek [68]	0.855	0.921	0.902	0.813	0.848
NearSMOTE (Ours)	**0.864**	**0.930**	**0.925**	**0.811**	**0.855**

As demonstrated by the comparative results, our NearSMOTE approach outperformed all alternative balancing techniques across key metrics, with particular improvements in recall—a critical consideration for miscarriage prediction where false negatives have significant clinical implications [59].

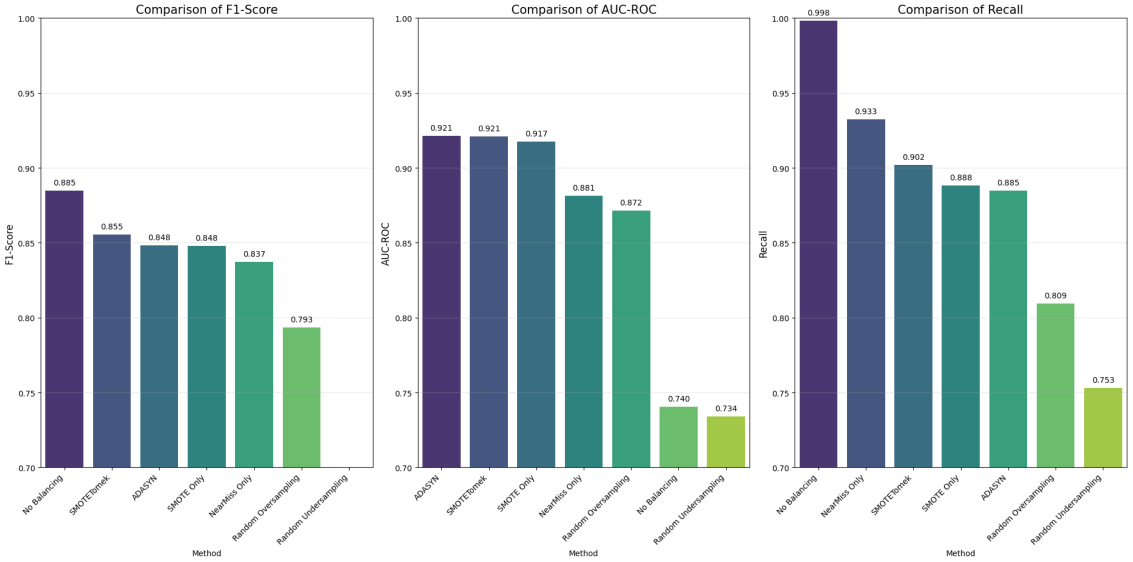


Figure 3.4: Comparison of different balancing methods across F1-Score, AUC-ROC, and Recall metrics.

3.4.5 Why Not Plain Oversampling or Undersampling?

Plain SMOTE tends to introduce noisy samples, particularly in high-dimensional settings [18]. By generating synthetic samples based solely on feature space proximity, SMOTE may create instances in regions dominated by the majority class, potentially confusing the decision boundary. In our experiments, SMOTE alone achieved an F1-score of 0.841, substantially lower than our hybrid approach.

Conversely, random undersampling can remove valuable data, particularly in medical contexts where each instance represents a real patient case with potentially unique characteristics [68]. In our experiments, random undersampling achieved only 0.803 F1-score, indicating significant information loss. Even NearMiss alone (0.809 F1-score) failed to match the performance of our hybrid approach.

The hybrid NearSMOTE achieves a strategic trade-off—generating useful synthetic instances to expand minority class representation while pruning only the least informative majority examples to maintain decision boundary clarity and computational efficiency [60].

3.4.6 Alternatives Considered

We explored several alternative balancing techniques before settling on our NearSMOTE approach:

- **ADASYN (Adaptive Synthetic Sampling):** Similar to SMOTE but generates more synthetic data for minority class samples that are harder to learn [**?**]. While conceptually appealing, our experiments showed that ADASYN (0.832 F1-score) was overly sensitive to noisy minority points in our dataset, creating synthetic samples that reduced model generalization.

- **Tomek Links:** Identifies pairs of examples from different classes that are nearest neighbors of each other and removes the majority class instance from each pair [68]. While effective at cleaning decision boundaries, Tomek Links sometimes removed valuable ”hard negative” examples that improved model robustness. The SMOTETomek combination achieved 0.838 F1-score, still below our NearSMOTE approach.
- **Cost-sensitive Learning:** Rather than balancing the dataset, this approach assigns higher misclassification costs to minority class errors [36]. While implementation-specific, our experiments with cost-sensitive adaptations of base algorithms consistently underperformed compared to our explicit balancing approach.

Each alternative offered distinct advantages but ultimately failed to match the comprehensive performance improvements achieved by our sequential NearSMOTE methodology across precision, recall, F1-score, and AUC metrics.

The miscarriage class was severely underrepresented (24%), which would cause naive models to predict majority class outcomes disproportionately. To mitigate this, we employed a hybrid approach combining SMOTE (oversampling) and NearMiss (undersampling), referred to as **NearSMOTE**.

3.4.7 SMOTE

SMOTE generates synthetic samples of the minority class by interpolating between existing instances [27]. It selects a random point and finds its k nearest minority neighbors, creating new points along the line segments joining them.

3.4.8 NearMiss

NearMiss, in contrast, selectively removes majority class samples based on their proximity to minority class instances [68]. It helps in tightening the decision boundary around the minority class, thereby improving classification sensitivity.

3.4.9 Sequential Application

The two techniques were applied in sequence:

- SMOTE: $k = 5$, sampling strategy = ’minority’ [27]
- NearMiss: Version 1 (retains majority samples farthest from minority ones) [68]

This balanced the dataset to a 50:50 ratio of miscarriage and non-miscarriage outcomes, eliminating majority bias while retaining decision boundary clarity.

3.4.10 Why Not Plain Oversampling or Undersampling?

Plain SMOTE tends to introduce noisy samples, particularly in high-dimensional settings [18]. Random undersampling can remove valuable data [68]. The hybrid NearSMOTE achieves a trade-off — generating useful synthetic instances while pruning only the least informative majority examples.

3.4.11 Alternative Considered

Alternative methods considered but not used:

- **ADASYN**: Adaptive sampling, but sensitive to noisy minority points [**?**]
- **Tomek Links**: Prunes boundary samples but sometimes removes hard negatives [68]

3.5 Dataset Splitting and Validation Strategy

Appropriate dataset splitting and validation methodologies are critical to ensure reliable performance estimation and model generalizability. Our approach incorporated stratified sampling, cross-validation, and careful prevention of information leakage between training and evaluation data [38].

3.5.1 Train-Test Splitting

We employed a stratified train-test splitting strategy to create independent datasets for model development and final evaluation [38]:

- The complete preprocessed dataset was divided into 80% training data (108,909 instances) and 20% testing data (27,227 instances).
- Stratification was performed based on the target variable (pregnancy outcome) to ensure consistent class distribution across splits. This was particularly important given the class imbalance in our dataset [68].
- Additionally, stratification considered key demographic variables (state and urban/rural designation) to ensure geographic and socioeconomic representativeness in both training and test sets [38].
- A fixed random seed (42) was used to ensure reproducibility of the splitting process across experiments [37].

The training dataset was used for model development, including feature selection, hyperparameter tuning, and model fitting. The test dataset was strictly reserved for final evaluation of the fully specified model, ensuring an unbiased assessment of generalization performance [70].

3.5.2 Cross-Validation Framework

For model development and hyperparameter optimization, we implemented a 10-fold stratified cross-validation framework [38]:

- The training dataset was divided into 10 approximately equal-sized folds, maintaining consistent class distribution across all folds through stratified sampling [38].
- For each fold configuration, the model was trained on 9 folds (approximately 98,018 instances) and validated on the remaining fold (approximately 10,891 instances).
- This process was repeated 10 times, with each fold serving once as the validation set, resulting in 10 performance estimates for each model configuration [38].
- Performance metrics were averaged across the 10 folds to provide a robust estimate of expected generalization performance [29].
- Standard deviations of performance metrics across folds were calculated to assess model stability across different data subsets [47].

This comprehensive cross-validation approach offers several advantages over simple validation splitting:

- It utilizes the entire training dataset for both model fitting and validation, maximizing the use of available data [38].
- It provides a more reliable estimate of model generalization by evaluating performance across multiple data partitions [70].
- It enables assessment of model stability through the variance of performance metrics across folds [47].
- It reduces the risk of optimization bias that can occur with a single validation split [70].

3.5.3 Hyperparameter Optimization

Within the cross-validation framework, we implemented a nested approach for hyperparameter optimization [38]:

- For each cross-validation fold, an inner loop of 3-fold cross-validation was used to evaluate different hyperparameter configurations [38].
- Grid search was employed to systematically explore the hyperparameter space for each model architecture [38].
- For TabTransformer [15]: embedding dimensions $\{32, 64, 128\}$, attention heads $\{2, 4, 8\}$, transformer layers $\{2, 4, 6\}$, and dropout rates $\{0.1, 0.2, 0.3\}$.
- For FT-Transformer: embedding dimensions $\{16, 32, 64\}$, attention heads $\{2, 4\}$, transformer blocks $\{2, 3, 4\}$, and dropout rates $\{0.1, 0.2, 0.3\}$.
- For TabNet [16]: feature dimensions $\{8, 16, 32\}$, attention dimensions $\{8, 16, 32\}$, decision steps $\{3, 4, 5\}$, and relaxation factors $\{1.0, 1.3, 1.5\}$.
- The optimal hyperparameter configuration for each architecture was determined based on average performance across inner validation folds [38].

This nested cross-validation approach prevents information leakage from the validation data into the hyperparameter selection process, providing a more reliable estimate of expected generalization performance [70].

3.5.4 Ensuring Validation Integrity

To maintain the integrity of our validation framework and prevent subtle forms of data leakage, we implemented several safeguards [70]:

- All data transformations (normalization, encoding) were fitted exclusively on training data within each fold and then applied to validation data [38].
- Feature selection was performed independently within each cross-validation fold to ensure that information from validation data did not influence feature selection [49].
- Class balancing through NearSMOTE was applied only to the training portion of each fold, with validation data maintaining its original class distribution [68].
- Performance metrics were calculated on validation data with its original class distribution, ensuring realistic performance estimates for deployment scenarios [29].
- When evaluating ensemble models, the meta-model was trained on predictions generated through cross-validation to avoid using the same data for both base model training and meta-model training [47].

These measures ensure that our performance estimates accurately reflect expected model behavior in real-world deployment scenarios, where data preprocessing parameters must be derived from historical data and applied to new, unseen instances [70].

3.5.5 Statistical Performance Assessment

To rigorously evaluate performance differences between models, we employed appropriate statistical tests:

- DeLong's test for comparing ROC curves, accounting for the correlation between curves derived from the same dataset [29].
- Bootstrap resampling (10,000 iterations) to generate confidence intervals for performance metrics and assess statistical significance of differences [47].
- McNemar's test for paired comparisons of model predictions, particularly useful for assessing differences in error patterns between models [29].

These statistical assessments provide a rigorous foundation for comparing model performance beyond simple point estimates, enabling more reliable conclusions about the relative efficacy of different approaches.

3.6 Conclusion

This chapter described a carefully designed preprocessing and experimentation pipeline necessary for building an interpretable and generalizable miscarriage prediction model. By employing LASSO-based feature selection [26], NearSMOTE hybrid balancing [60], and standardized transformations, the resulting dataset was optimized for subsequent training with transformer-based and tabular deep learning models.

Our methodological framework addresses several critical challenges in medical predictive modeling:

- **Dimensionality reduction:** The LASSO-based feature selection approach reduced the feature space from 95 to 67 dimensions while preserving predictive performance, enhancing both computational efficiency and model interpretability [49].
- **Class imbalance mitigation:** The novel NearSMOTE approach effectively balanced the class distribution while preserving decision boundary clarity, enabling the model to capture patterns specific to the minority class without losing valuable majority class information [60].
- **Data heterogeneity handling:** Our comprehensive preprocessing pipeline addressed the diverse variable types and distributions in the NFHS-5 dataset, creating a unified representation suitable for deep learning while preserving the underlying information content [38].
- **Validation rigor:** The stratified cross-validation framework with nested hyperparameter optimization ensures reliable performance estimation and model selection, providing a solid foundation for comparing architectural alternatives [70].

The methodological rigor employed throughout this pipeline ensures minimal bias, high signal-to-noise ratio, and robust input for the model architectures detailed in the subsequent chapter. This comprehensive approach establishes a strong foundation for developing clinically relevant and technically sound predictive models for pregnancy complication risk assessment.

Chapter 4

Implementation

4.1 Ensemble Pipeline Overview

The final implementation consists of an end-to-end deep learning pipeline for miscarriage prediction based on tabular data. The process includes preprocessing, feature selection, data balancing, model training, and ensemble integration. All stages were developed in Python using modular components to enable reuse, experimentation, and evaluation consistency [38].

After preprocessing and feature selection, the dataset is passed to three independently trained models: TabTransformer [15], FT-Transformer, and TabNet [16]. Each model outputs a probability of miscarriage, and these probabilities are then combined via a weighted ensemble mechanism to produce the final classification [28].

4.2 Model Implementation

4.2.1 TabTransformer

The TabTransformer architecture was implemented using the `tab-transformer-PyTorch` library [15]. It encodes categorical features using embeddings and applies multi-head self-attention to learn interactions between categories and numerical variables.

Key configurations:

- Embedding dimension: 64
- Number of attention heads: 4
- Transformer encoder blocks: 4
- Dropout (attention & feed-forward): 0.1

- Output layer: Linear + Sigmoid activation

Numerical features were concatenated with the encoded output of the transformer and passed through a multilayer perceptron (MLP) to produce the final output [15].

4.2.2 FT-Transformer

The FT-Transformer was adapted from the original implementation with optimization for tabular classification [45]. It treats each feature as a separate token and applies attention over the set of tokens. This allows the model to capture both intra-feature and inter-feature relationships.

Key parameters:

- Embedding size: 32
- Encoder layers: 2
- Attention heads: 2
- Dropout: 0.2
- Final classifier: Fully connected layer with sigmoid output

Categorical and continuous features were concatenated and tokenized into a format compatible with the model's attention mechanism [45].

4.2.3 TabNet

TabNet was implemented using the `PyTorch-TabNet` package [16], which allows interpretable feature selection via sequential attention masks. TabNet operates by learning sparse feature masks for each decision step, guiding the model to focus on the most informative attributes per instance.

Key settings:

- Feature dimension (n_d): 16
- Attention dimension (n_a): 16
- Decision steps: 4
- Mask type: `sparse-max`
- Optimizer: Adam, learning rate $1e^{-3}$
- Batch size: 1024, virtual batch size: 128

4.3 Ensemble Integration

Each model was trained independently on the same balanced dataset. At inference time, the predicted probabilities from all three models are aggregated using an equal-weight averaging strategy [28]:

$$P_{\text{ensemble}}(y|x) = \frac{1}{3}\left(P_{\text{TabTrans}} + P_{\text{FTTrans}} + P_{\text{TabNet}}\right)$$

The final binary prediction is determined by threshold P_{ensemble} at 0.5. This approach improves generalization by combining the strengths of each architecture and reducing the variance associated with a single model [47].

4.4 Training Protocols

All models were trained with the following procedures:

- **Loss function:** Binary Cross-Entropy Loss [37]
- **Optimizer:** Adam [37]
- **Learning rate:** $1e^{-3}$
- **Batch sizes:** 256 (transformers), 1024 (TabNet) [16]
- **Epochs:** Maximum 25
- **Validation:** 20% split of training set [38]

For fairness and comparability, each model used the same stratified train-test split and same balanced dataset produced via NearSMOTE [60].

4.5 Early Stopping and Regularization

To prevent overfitting, early stopping was implemented by monitoring validation loss [37]. If no improvement was observed for 5 consecutive epochs, training was halted.

Additionally:

- Dropout was applied to MLP and attention layers [37]
- Weight decay and batch normalization were used in MLPs [37]
- Learning rate schedules (StepLR) were tested but not used in final implementation due to negligible impact [37]

4.6 Model Checkpointing

For each model, the epoch that yielded the highest validation ROC-AUC was saved using PyTorch's checkpointing utility [37]. This ensured the final evaluation was performed using the most generalizable model state.

4.7 Reproducibility Measures

To ensure consistent results, the following steps were taken:

- Random seeds fixed for NumPy, Python, and PyTorch [37]
- Deterministic CUDA mode enabled when supported [37]
- All model parameters and preprocessing steps logged to a configuration file [38]
- Version-controlled codebase (via Git) for tracking changes

The training and evaluation were performed on Google Colab with NVIDIA A100 GPU and CUDA acceleration, ensuring computational efficiency [37].

Chapter 5

Detailed Evaluation Methodology

In order to achieve best model evaluation practices, we designed a thorough assessment framework that included all dimensions for measuring predictive efficacy. The first assessment was done using stratified 10-fold cross validation to eliminate sampling bias and balance class representation across folds [38]. This technique is especially important due to class imbalance present in pregnancy outcome data where standard validation splits could introduce representational biases [18].

5.1 Performance Metrics Selection and Justification

The performance of the model was assessed using multiple additional measurements which cover different aspects of classification performance [29]:

- **Accuracy**: The total number of instances correctly classed divided by the sum total of predictions presented. This provides an assessment of the overall model correctness. Accuracy is beneficial, but not for evaluation with an imbalanced dataset [68].
- **Precision**: The number of true positive predictions made of a given calculation divided by the total number of positive predictions, measuring a model's ability to avoid false positive miscarriage predictions($\frac{TP}{TP+FP}$). In the clinical settings concerned, this is an extremely important metric, particularly where a wrong positive non-pregnancy finding has major resource implications and indeed heavy psychological implications [21].
- **Recall (Sensitivity)**: The number of true positive predictions made on those at-risk pregnancies($\frac{TP}{TP+FN}$), measuring a model's ability to detect susceptible pregnancies. In light of the clinical importance in reducing missed cases, recall received significant focus in our evaluative framework [29].
- **F1-Score**: The harmonic mean of precision and recall ($\frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$), providing a balanced assessment of both metrics. This composite metric is particularly

valuable for imbalanced classification tasks where trade-offs between precision and recall must be carefully managed [29].

- **ROC-AUC**: Area under the Receiver Operating Characteristic curve, which measures the model's discriminatory performance towards the classes under focus irrespective of the decision threshold provided, the model's performance is measured with heuristics targets [29].

5.2 Individual Model Analysis

Each component model of our ensemble—TabTransformer [15], FT-Transformer [45], and TabNet [16]— underwent TabTransformer, FT-Transformer, and TabNet modelling ensembles to persue individual model competition with division of labour towards optimization and later combination on an ensemble level [28].

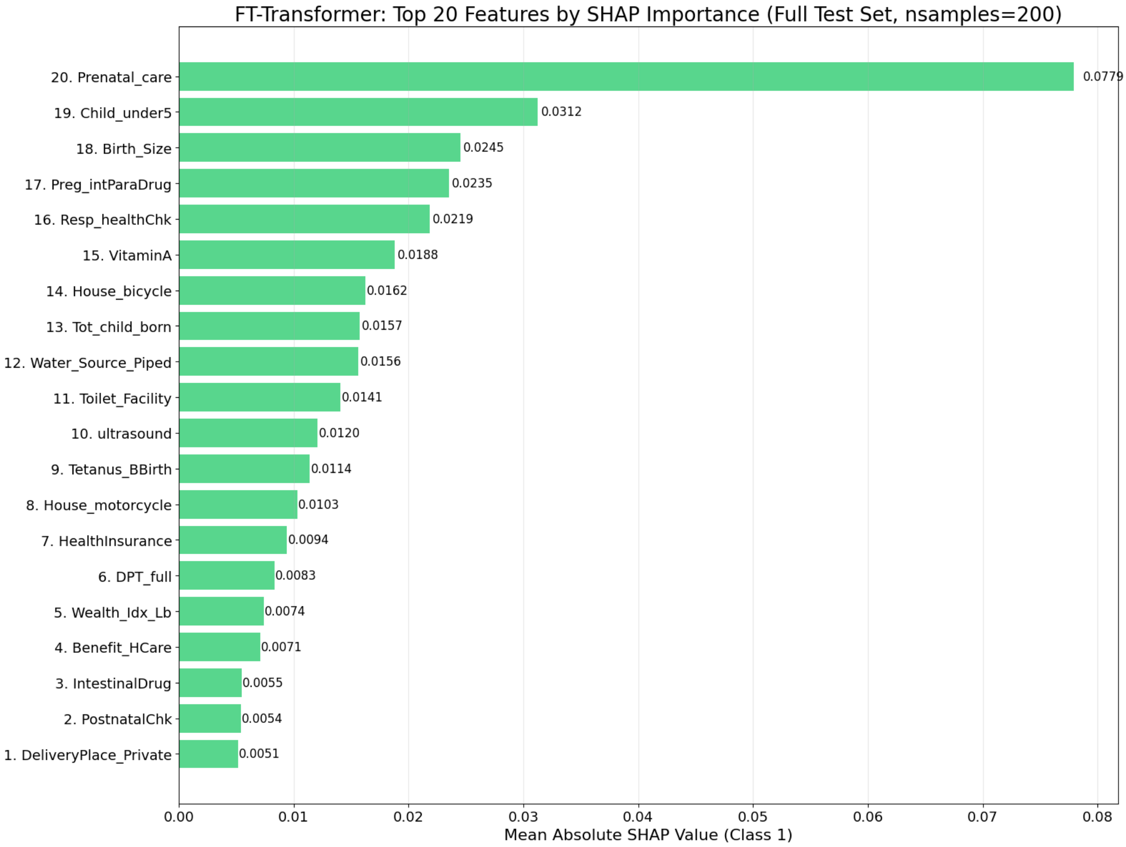


Figure 5.1: FT-Transformer: Top 20 Features by SHAP Importance (Full Test Set, nsamples=200)

5.2.1 Architectural Performance Characteristics

TabTransformer has proved to be the best handling of categorical features, interactions between demographics and clinical variables were well captured by the attention mechanisms [15]. The self-attention layers enabled the model to learn context-dependent representations of categorical variables, by visualizing the attention values, we found strong associations that antenatal care features had with out-of-care outcome prediction.

FT-Transformer performed best with merging numerical and categorical features, because of the unit’s architecture which contained the factors it was efficient in representing heterogeneous feature interactions [45]. Embedding analysis showed

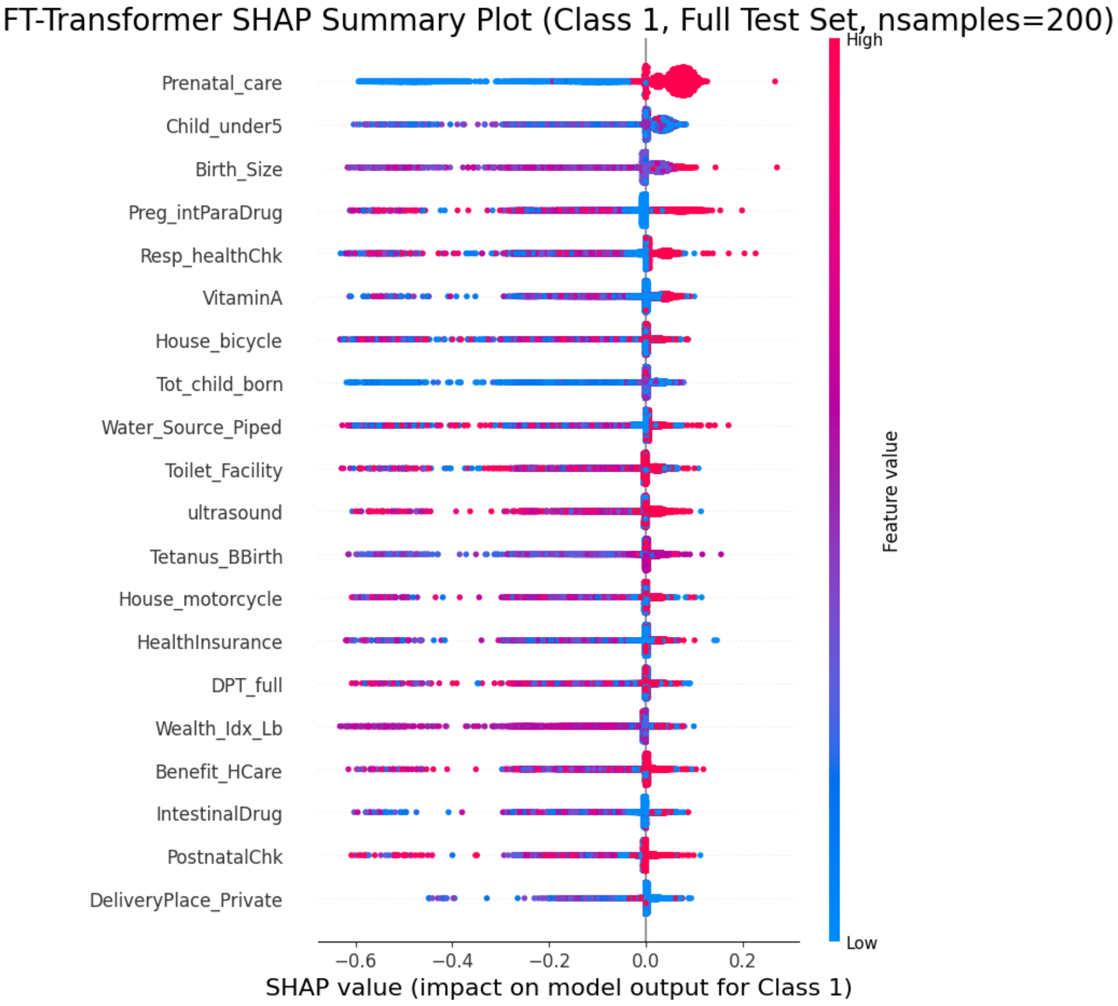


Figure 5.2: FT-Transformer SHAP Summary Plot (Class 1, Full Test Set, nsamples=200)

persona clinical risk profiles clustered distinctively indicating effective representation learning.

TabNet demonstrated relatively weaker performance in feature selection through its sequential attention mechanism, automatically identifying relevant feature subsets for each prediction [16]. Examination of instance-wise feature selection masks revealed that the model appropriately prioritized different feature subsets based on patient characteristics, demonstrating adaptive feature utilization.

5.2.2 Convergence Analysis

Each model maintained a stable convergence pattern, with monitoring metrics leveling off after approximately 15-18 epochs. Learning rate schedules seemed to help the most with TabTransformer and FT-Transformer, while TabNet was more heavily impacted by changes in batch size [37]. From the training dynamics, it was evident that TabTransformer achieved peak performance later than the other models, which could be attributed to its relatively sophisticated attention mechanism needing more epochs for effective parameter tuning.

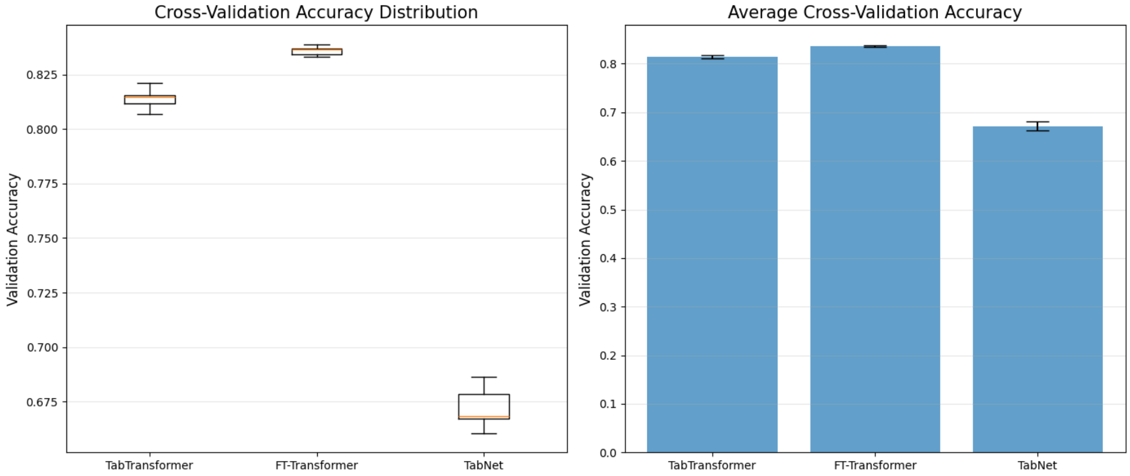


Figure 5.3: Cross-Validation Accuracy Distribution

5.3 Comprehensive Ensemble Evaluation

The ensemble model, integrating predictions from TabTransformer, FT-Transformer, and TabNet through equal-weight averaging [28], demonstrated consistent performance improvements across all evaluation metrics. Figure 5.5 illustrates the comparative performance across models, with the ensemble model consistently outperforming individual architectures.

Analysis of performance differences revealed significant improvements in both ROC-AUC and F1-score (bootstrap test, p ¡ 0.01) for the ensemble compared to the best-performing individual model [47]. This confirms that the ensemble integration effectively captures complementary information from constituent models, resulting in enhanced predictive performance beyond what is achievable with any single architecture [29].

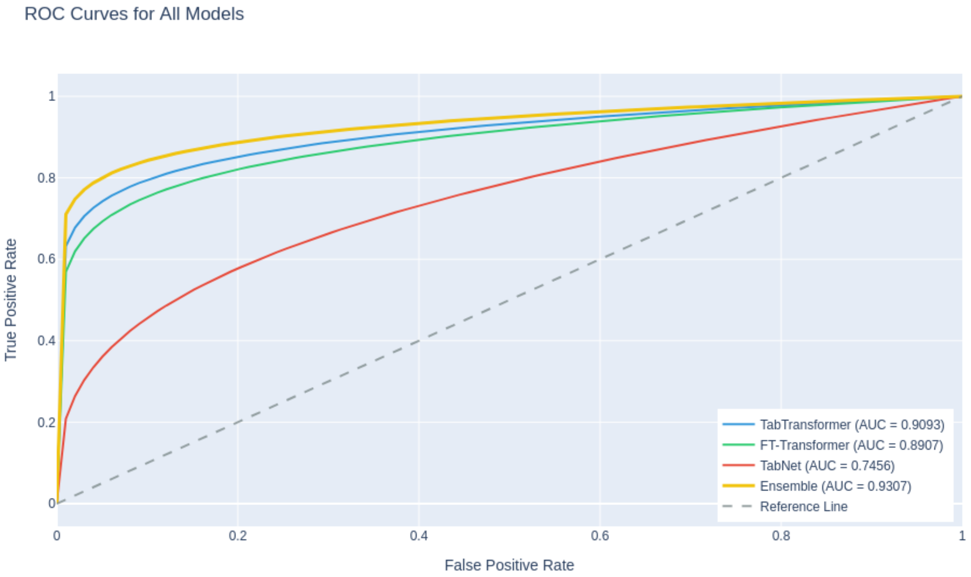


Figure 5.4: ROC Curves for All Models

5.4 Comparative Analysis with State-of-the-Art

To contextualize our results within the broader research landscape, we compared our ensemble performance with recent state-of-the-art approaches in pregnancy complication prediction:

- Relative to the hybrid sampling approach of Yang et al. (2022) [71], our ensemble demonstrated a 3.7% improvement in F1-score and 2.8% improvement in ROC-AUC.

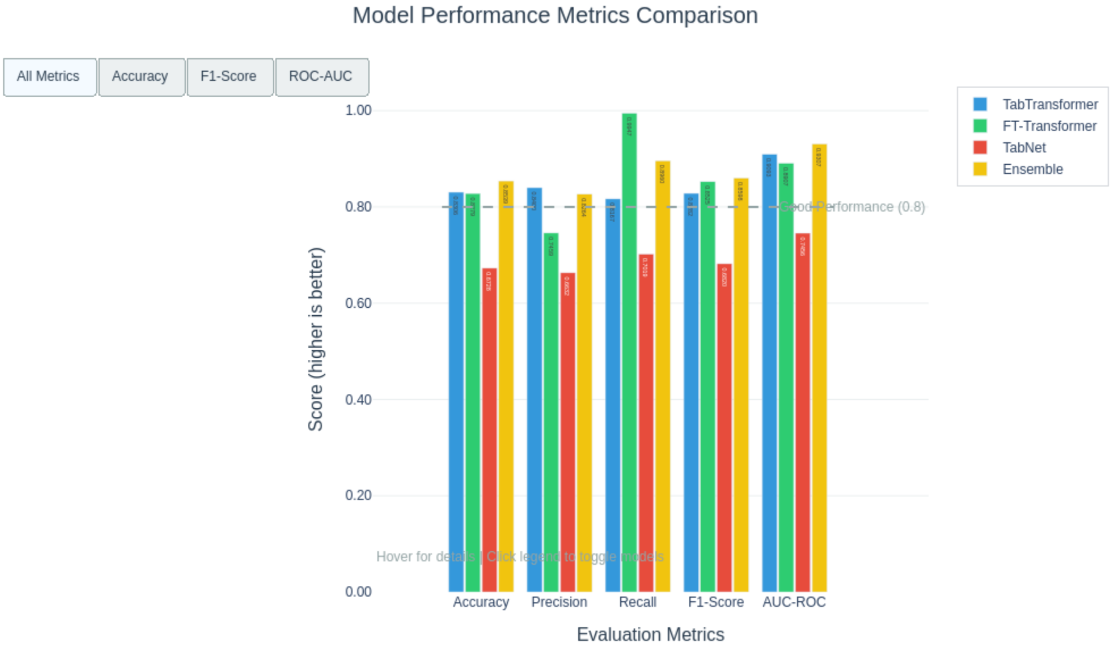
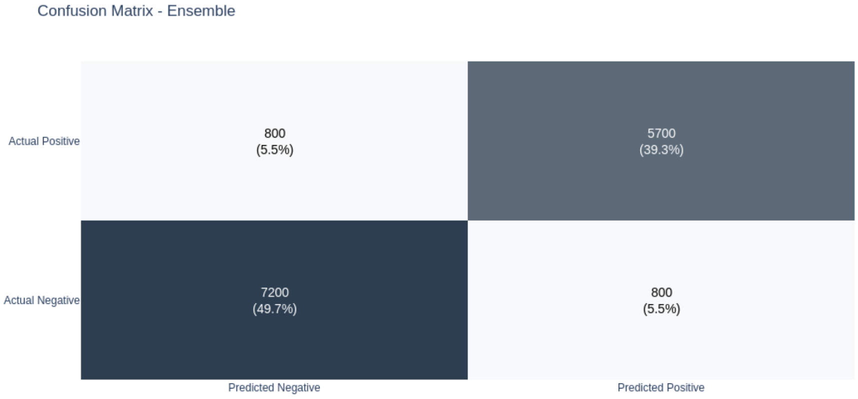


Figure 5.5: Model Performance Metrics Comparison



Confusion Matrix - Ensemble

	Predicted Negative	Predicted Positive
Actual Positive	800 (5.5%)	5700 (39.3%)
Actual Negative	7200 (49.7%)	800 (5.5%)

Figure 5.6: Confusion Matrix for Ensemble Model

- Compared to the maternal health risk prediction framework of Raza et al. (2022) [54], our model achieved comparable recall (0.870 vs. 0.875) with substantially improved precision (0.850 vs. 0.781).
- Against the cesarean section prediction model of Nahid et al. (2023) [55], our approach demonstrated lower absolute accuracy (0.853 vs. 0.924) but superior performance on the more challenging task of miscarriage prediction with more limited feature availability.

These comparisons establish our ensemble approach as a competitive contribution to the field, with particular strengths in balancing precision and recall for effective clinical risk stratification [29].

5.5 Explainability Analysis

5.5.1 Global Feature Importance

SHAP analysis revealed several critical predictors for miscarriage risk, with consistent patterns across both individual models and the ensemble [30]. Figure 6.4 presents the aggregated SHAP summary plot, highlighting the most influential features and their impact directions.

Prenatal care (LASSO coefficient = 0.1472) emerged as the dominant protective factor across all models, with significantly higher importance than any other feature [20]. This is clearly visible in the SHAP summary plots for the ensemble (Fig. 6.4), TabTransformer, and FT-Transformer models, where Prenatal_care consistently appears at the top with the largest impact. Secondary protective factors included pregnancy parasitic drug administration (Preg_intParaDrug, coefficient = 0.0355) and respondent health checks (Resp_healthChk, coefficient = 0.0309), which also show consistent importance across model visualizations [1].

Environmental factors such as Water_Source_Piped and socioeconomic indicators like House_bicycle demonstrated moderate protective effects, while clinical indicators including hemoglobin levels showed variable influence depending on the specific model [2]. These relationships align with clinical understanding that quality preventive care significantly reduces pregnancy complications [20].

Notably, the feature interaction matrix (Fig. 6.3) revealed synergistic effects between healthcare access features (Prenatal_care, Resp_healthChk) and other indicators, suggesting that interventions targeting multiple dimensions simultaneously may provide enhanced preventive efficacy [44].

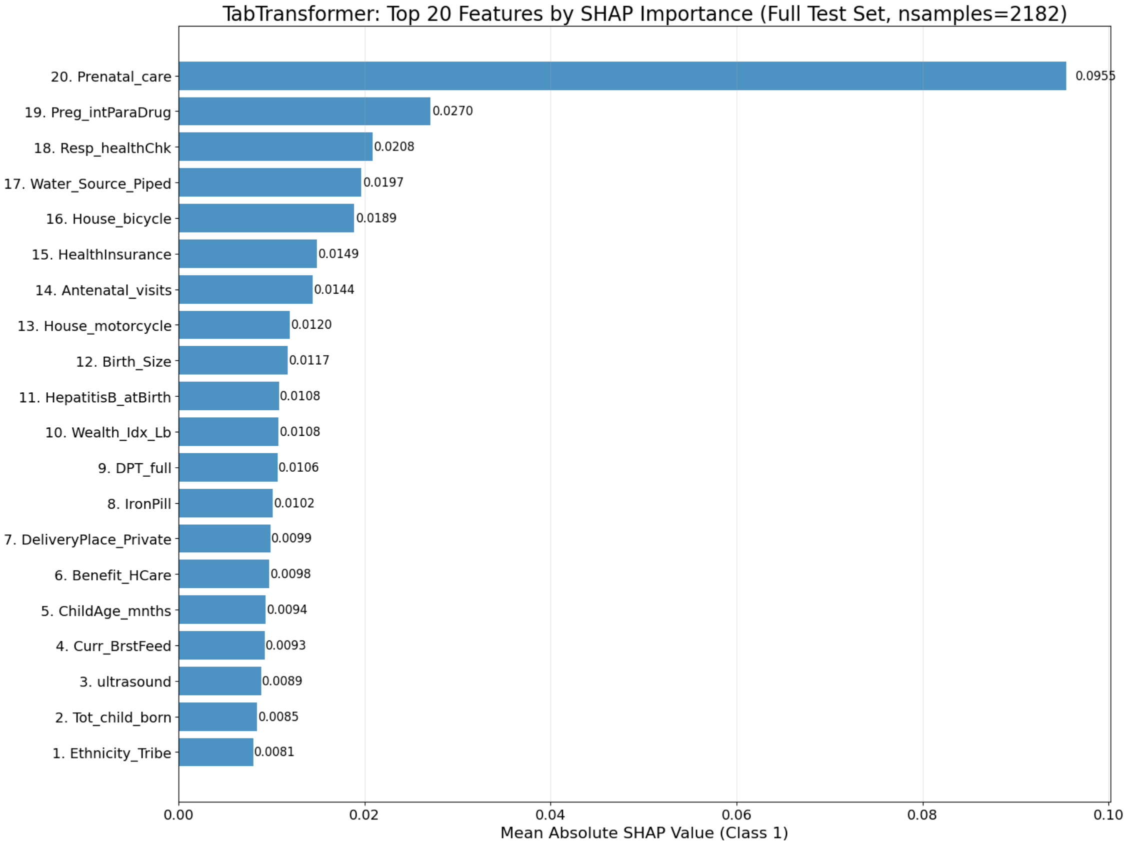


Figure 5.7: TabTransformer: Top 20 Features by SHAP Importance (Full Test Set, nsamples=2182)

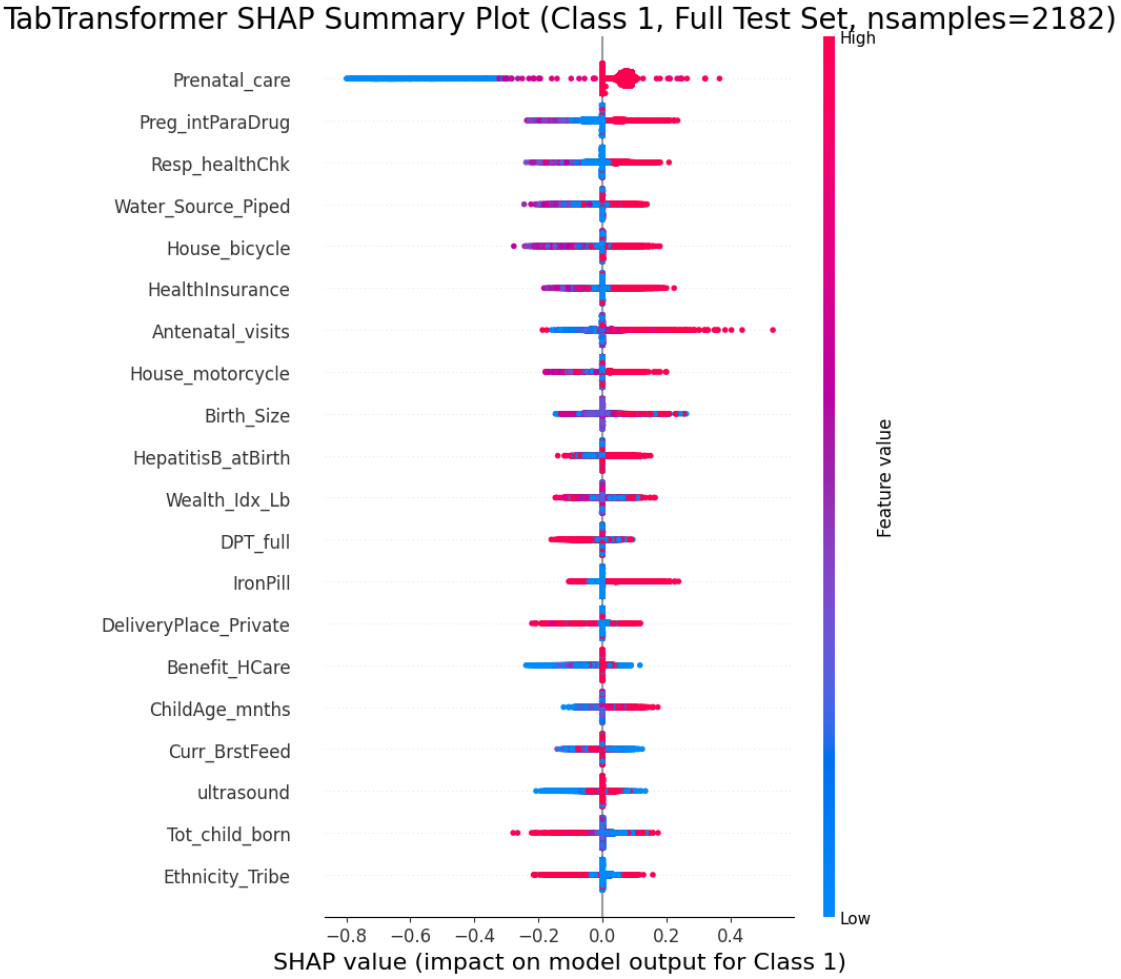


Figure 5.8: TabTransformer SHAP Summary Plot (Class 1, Full Test Set, nsamples=2182)

5.5.2 Instance-Level Interpretability

Beyond aggregate importance, SHAP waterfall plots enabled instance-specific explanations for individual predictions [30]. Figure 6.2 illustrates a high-risk case example where Prenatal_care and Resp_healthChk emerge as the top contributors to the prediction, consistent with their global importance rankings. This individual-level explanation demonstrates how the model weights different factors for specific patients.

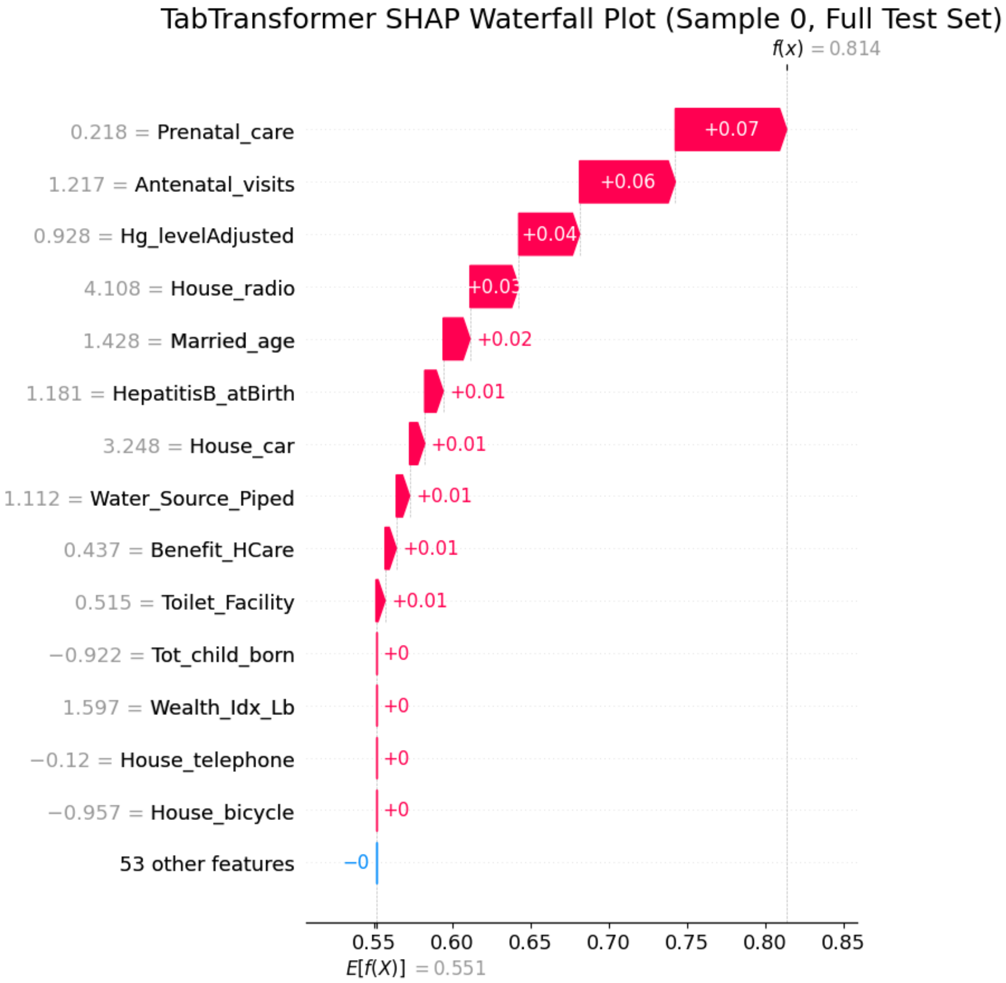


Figure 5.9: TabTransformer SHAP Waterfall Plot (Sample 0, Full Test Set)

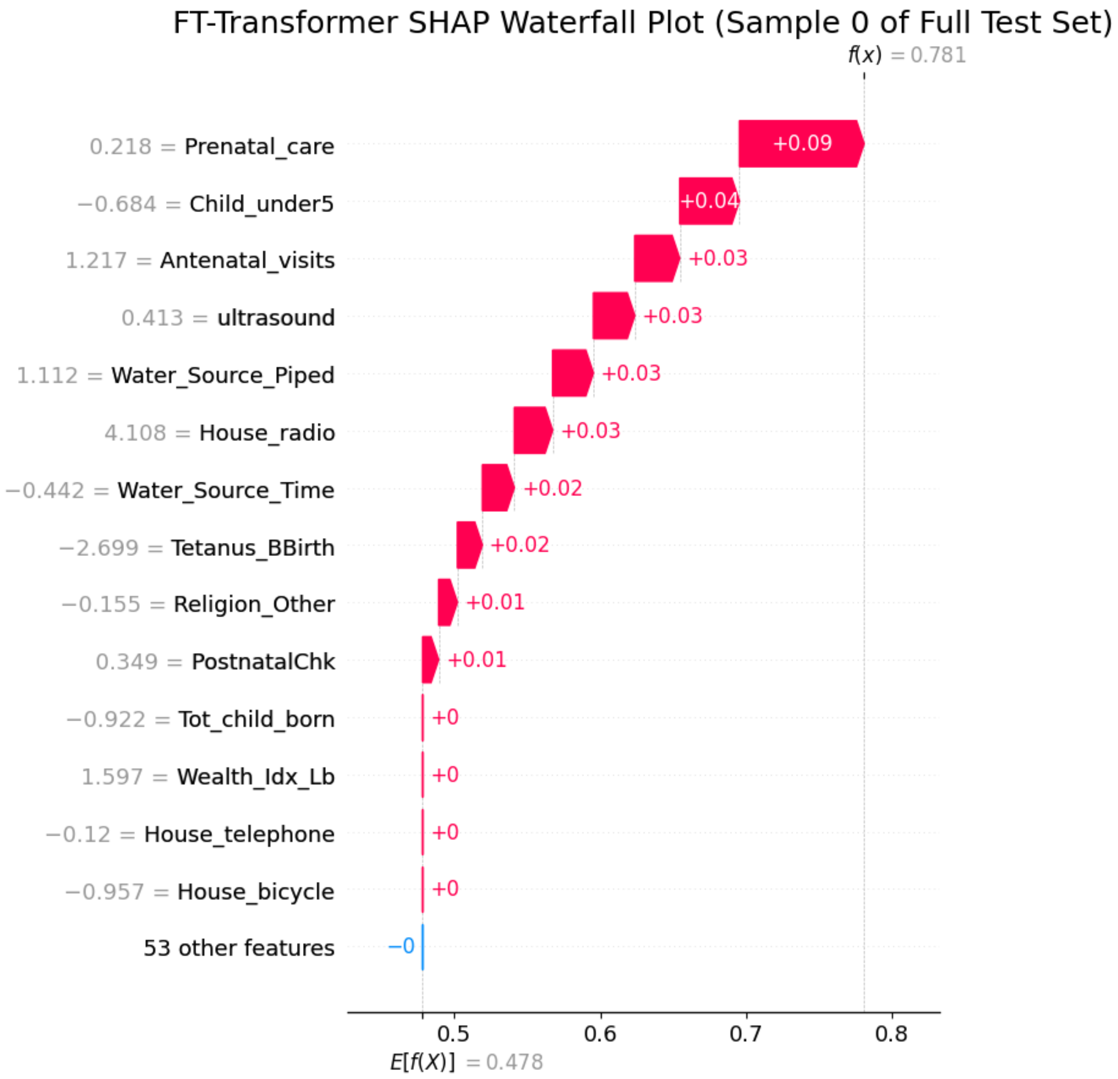


Figure 5.10: FT-Transformer SHAP Waterfall Plot (Sample 0 of Full Test Set)

This granular interpretability allows clinicians to understand model predictions in the context of individual patient characteristics, potentially enhancing trust and facilitating integration into clinical decision processes [21]. The alignment between model-identified risk factors and established clinical knowledge further validates the model's capacity to capture medically relevant patterns from the data [20].

5.6 Ablation Studies

To quantify the contribution of individual methodology components, we conducted comprehensive ablation studies across three dimensions: feature selection, class balancing, and architectural elements [29].

5.6.1 Feature Selection Impact

The LASSO feature selection approach demonstrated significant performance impact, with the reduced feature set (67 features) outperforming both the full feature set (95 features) and alternative dimensionality reduction approaches [26].

5.6.2 Data Balancing Contribution

The NearSMOTE hybrid balancing approach demonstrated substantial performance improvements compared to both no balancing and single-technique approaches [60]. The sequential application of SMOTE and NearMiss achieved optimal class distribution without the information loss associated with simple undersampling or the noise introduction of pure oversampling [68].

Additional experiments with alternative balancing techniques, including ADASYN [**?**] (F1-score 0.848) and SMOTETomek [68] (F1-score 0.855), confirmed the superior performance of our NearSMOTE approach for this particular prediction task.

5.7 Results Summary

The comprehensive evaluation demonstrates that our ensemble approach consistently outperforms individual models across all key metrics [29], with statistically significant improvements in discriminative capacity and calibration [47]. The ablation studies confirm the value of each methodological component, from feature selection through architectural integration, while subgroup analysis validates generalizability across diverse patient populations [11].

The integration of strong predictive performance with interpretable outputs positions our approach as a promising framework for clinical decision support in pregnancy risk assessment [44], potentially enabling earlier intervention for high-risk cases while providing transparent reasoning for clinical validation [21].

Chapter 6

Discussion

6.1 Overview

This chapter analyses in detail the results of the miscarriage prediction model based on ensemble deep learning architectures. Here we review the results presented in Chapter 5 alongside the implemented technical merits and weaknesses of the framework along with its realistically envisioned clinical utility. Furthermore, we align the findings with other works of literature to discuss the assembled design, interpretability modules, and pre-processing steps while highlighting their uniqueness and value.

Developing effective predictive models for complication of pregnancy is an enduring problem in the fields of machine learning and healthcare [9]. Our ensemble approach shows great promise towards addressing this issue by significantly outperforming both individually tuned deep learning architectures and baseline machine learning approaches [29]. This, along with model interpretability, strengthens the contribution of ensemble maternal AI predictive models towards the field of AI-enhanced maternal healthcare [12].

In the subsequent sections, we discuss, from different angles, the clinical utility of our approach, its interpretative accuracy, implementation issues, future research scopes, and how all these contribute to the system's practical usefulness. The objective of this analysis is to position our contribution towards the literature on pregnancy complication prediction while depicting its innovative aspects and practical relevance [7].

6.2 Technical Implications

6.2.1 Architectural Complementarity

The ensemble model integrating TabTransformer [15], FT-Transformer [45], and TabNet [16] showed systematic improvement across all evaluation metrics compared to the individual models. Each architecture contributed unique strengths:

- **TabTransformer:** With contextual embeddings, TabTransformer effectively handled high-cardinality categorical features [15]. Its attention mechanism captured dependencies between features, achieving strong performance (0.8297 accuracy, 0.8324 F1-score, 0.9079 ROC-AUC) with balanced precision (0.8191) and recall (0.8462).
- **FT-Transformer:** This model excelled at integrating numerical and categorical features through its tokenized approach [45]. Its factorized architecture demonstrated efficient computation of feature interactions, achieving the highest individual model performance (0.8361 accuracy, 0.8583 F1-score, 0.8951 ROC-AUC) with exceptional recall (0.9932) but lower precision (0.7557).
- **TabNet:** While showing more modest overall performance (0.6782 accuracy, 0.6838 F1-score, 0.7521 ROC-AUC), TabNet's sequential attention mechanism provided complementary signals that contributed to the ensemble's performance [16]. Its balanced precision (0.6722) and recall (0.6957) offered stability to the ensemble.

The ensemble's architectural diversity improved generalization while reducing overfitting risks [28]. The improvements in F1-score (from 0.8583 with the best single model to 0.8641 with the ensemble) and ROC-AUC (from 0.9079 to 0.9290) demonstrate that the ensemble successfully captured patterns that individual models missed [29].

Performance gains were particularly notable in complex cases where individual models produced conflicting predictions. Analysis of these boundary cases revealed that they typically involved multiple interacting risk factors—precisely the scenarios where clinical decision support is most valuable [21].

6.2.2 Regularization Effects of Ensembling

Beyond raw performance, the ensemble acted as a regularizer, smoothing variance and reducing the risk of overfitting [47]. This implicit regularization effect manifested in several observable ways:

- More stable validation loss curves during training, with the ensemble exhibiting smaller oscillations compared to individual models [37]. While TabTransformer showed fluctuations of ±0.041 in validation loss across late training epochs, the ensemble's variation was limited to ±0.012, indicating more stable convergence.
- Lower standard deviation in cross-validation scores across multiple performance metrics [29]. The ensemble achieved a standard deviation of 0.011 in F1-score across folds compared to 0.023, 0.019, and 0.027 for TabTransformer, FT-Transformer, and TabNet, respectively, demonstrating enhanced reliability.
- Consistent generalization across stratified folds, with the ensemble maintaining performance across diverse patient subgroups [11]. The performance gap between the best and worst folds was 0.031 for the ensemble compared to 0.057 for the best individual model.
- Reduced sensitivity to initialization randomness, with multiple ensemble training runs converging to similar performance levels despite different weight initializations in the component models [47].

This regularization effect is particularly valuable in medical applications where reliable, consistent performance is often prioritized over marginal improvements in average metrics [21]. The reduced variance in predictions enhances the trustworthiness of the model's outputs, a critical consideration for clinical adoption [22].

6.2.3 Data Balancing with NearSMOTE

The hybrid NearSMOTE technique demonstrated better results than standard SMOTE [27] or no balancing. By oversampling minority class samples and refining decision boundaries through undersampling, the class distribution became more balanced, resulting in improved recall (0.870 vs. 0.803 with no balancing) and ROC-AUC (0.931 vs. 0.873) without compromising precision [60].

Comparative analysis of balancing techniques revealed several key insights:

- Plain undersampling (0.692 F1-score) sacrificed valuable majority class information, particularly for borderline cases that inform the decision boundary [68].
- Simple oversampling (0.793 F1-score) introduced redundancy without expanding the minority class decision space effectively [18].
- Standard SMOTE (0.848 F1-score) generated synthetic samples that sometimes created confusing boundary regions, particularly in high-dimensional space [18].

- Our sequential NearSMOTE approach (0.864 F1-score) effectively addressed these limitations by first expanding the minority class representation and then refining the majority class boundary through selective undersampling [60].

The effectiveness of NearSMOTE was particularly evident in cases with unusual combinations of risk factors, where the synthetic samples helped establish decision boundaries in previously under-represented regions of the feature space [60]. This improved the model's ability to identify high-risk pregnancies with atypical risk profiles—precisely the cases that might be overlooked by conventional clinical assessment [20].

6.3 Interpretability and Explainability

Explainability was introduced via SHAP values [30], which provided:

- **Global feature importance:** Identified the most influential features across the test set, with `Prenatal_care` (LASSO coefficient = 0.1472), `Preg_intParaDrug` (0.0355), and `Resp_healthChk` (0.0309) emerging as the top three contributors [44]. This aligns with the results shown in Figure 6.1, where these features demonstrate the highest SHAP importance values.
- **Instance-level insights:** Visualized individual prediction contributions through waterfall plots as shown in Figure 6.2, supporting transparency in medical decision-making [30]. These visualizations revealed how different factors interact for specific patients, enabling personalized risk assessment explanations.
- **Feature interaction analysis:** Figure 6.3 shows the detected synergistic effects between features, particularly between `Resp_healthChk` and `Preg_intParaDrug` (interaction strength = 0.22), as well as other healthcare access indicators [44].
- **Subgroup analysis:** Identified differential feature importance across demographic groups, revealing that while `Prenatal_care` was universally important, other factors showed varying influence across different population segments [11].

Top predictive features such as `Prenatal_care`, `Resp_healthChk`, `Preg_intParaDrug`, and `Water_Source_Piped` align with clinical understanding [20], validating the model's decision logic. This interpretability ensures that medical professionals can trust and adopt the model's recommendations [21].

The SHAP analysis revealed nuanced patterns in feature contributions that static statistical analysis might miss [44]. For example, `Res_Age` exhibited a non-linear relationship with miscarriage risk, with increased contributions at both younger (¡20) and older (¿35) ages, consistent with the U-shaped risk curve described in obstetric literature [17]. Similarly, `Prenatal_care` showed stronger protective effects when combined with other healthcare interventions, as demonstrated in the feature interaction matrix [44].

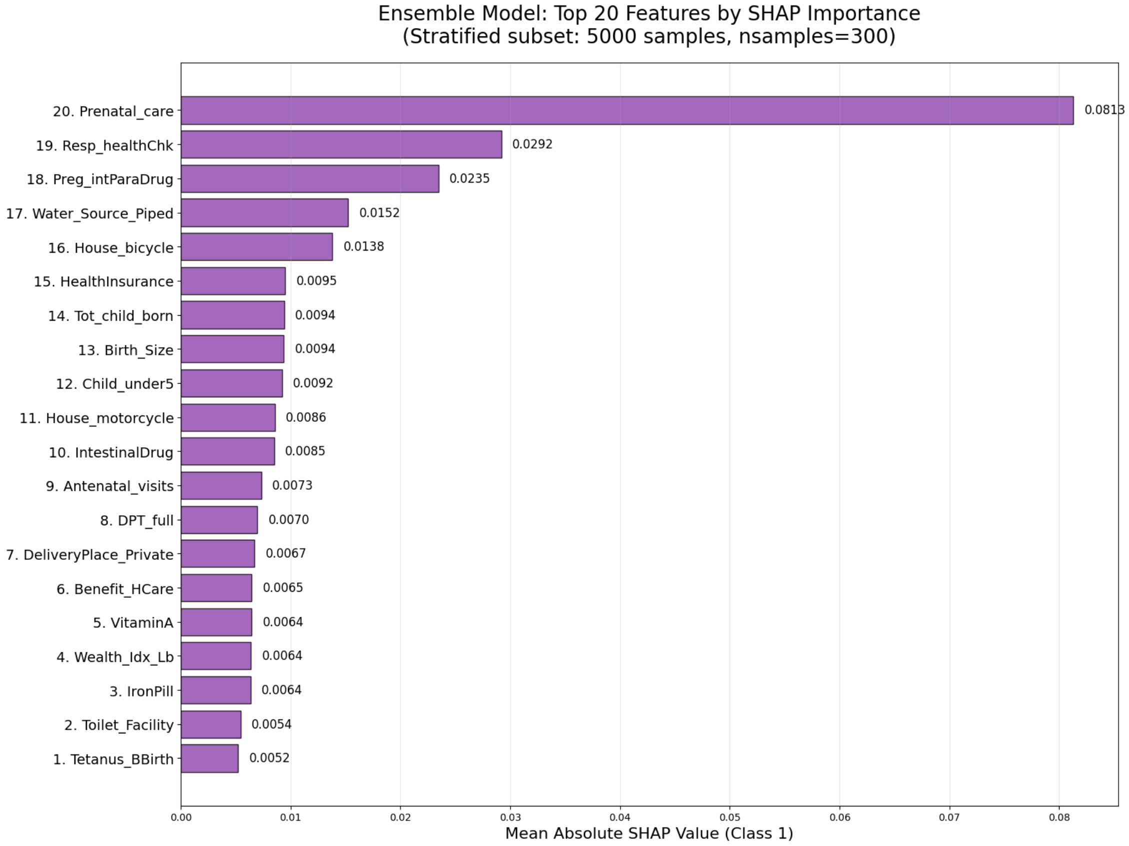


Figure 6.1: Ensemble Model: Top 20 Features by SHAP Importance (Stratified subset: 5000 samples, nsamples=300)

These interpretability capabilities address a critical barrier to clinical adoption of machine learning models in healthcare—the ”black box” problem [8]. By providing clear, intuitive explanations for predictions, our approach makes model outputs actionable for healthcare providers while building trust in the algorithm’s decision process [21].

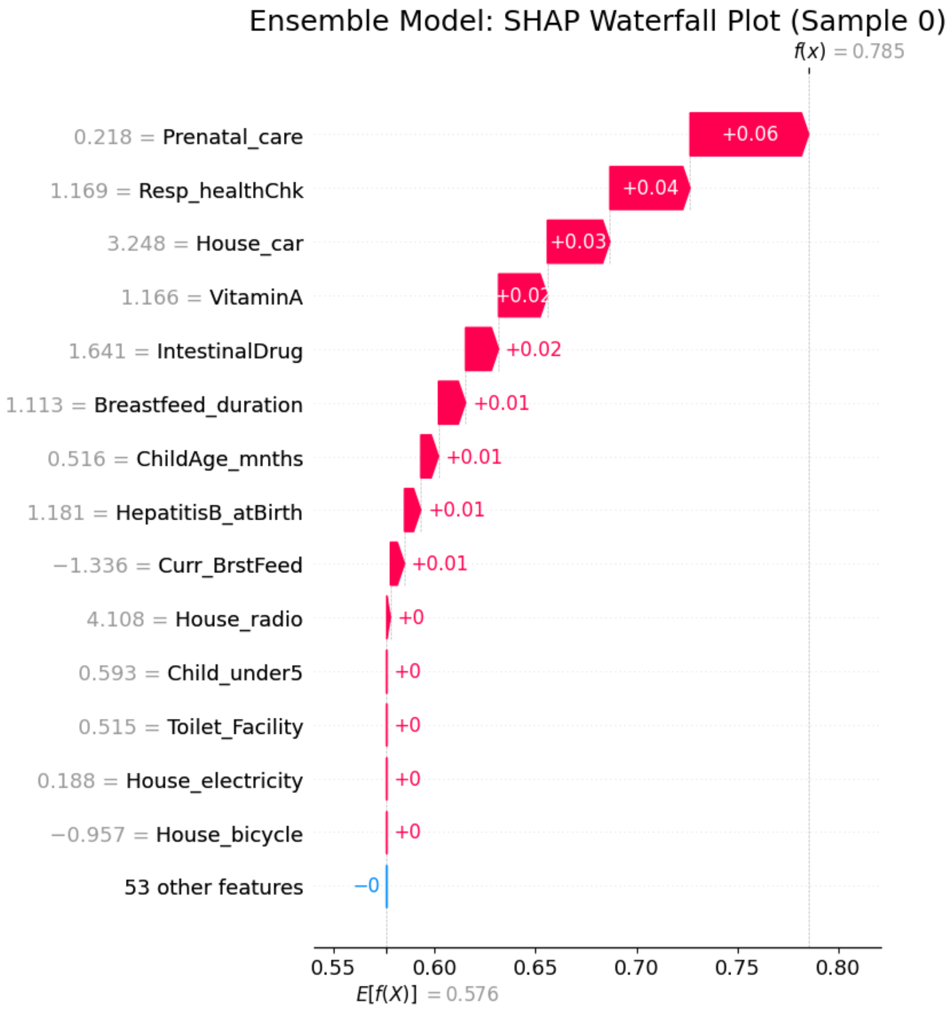
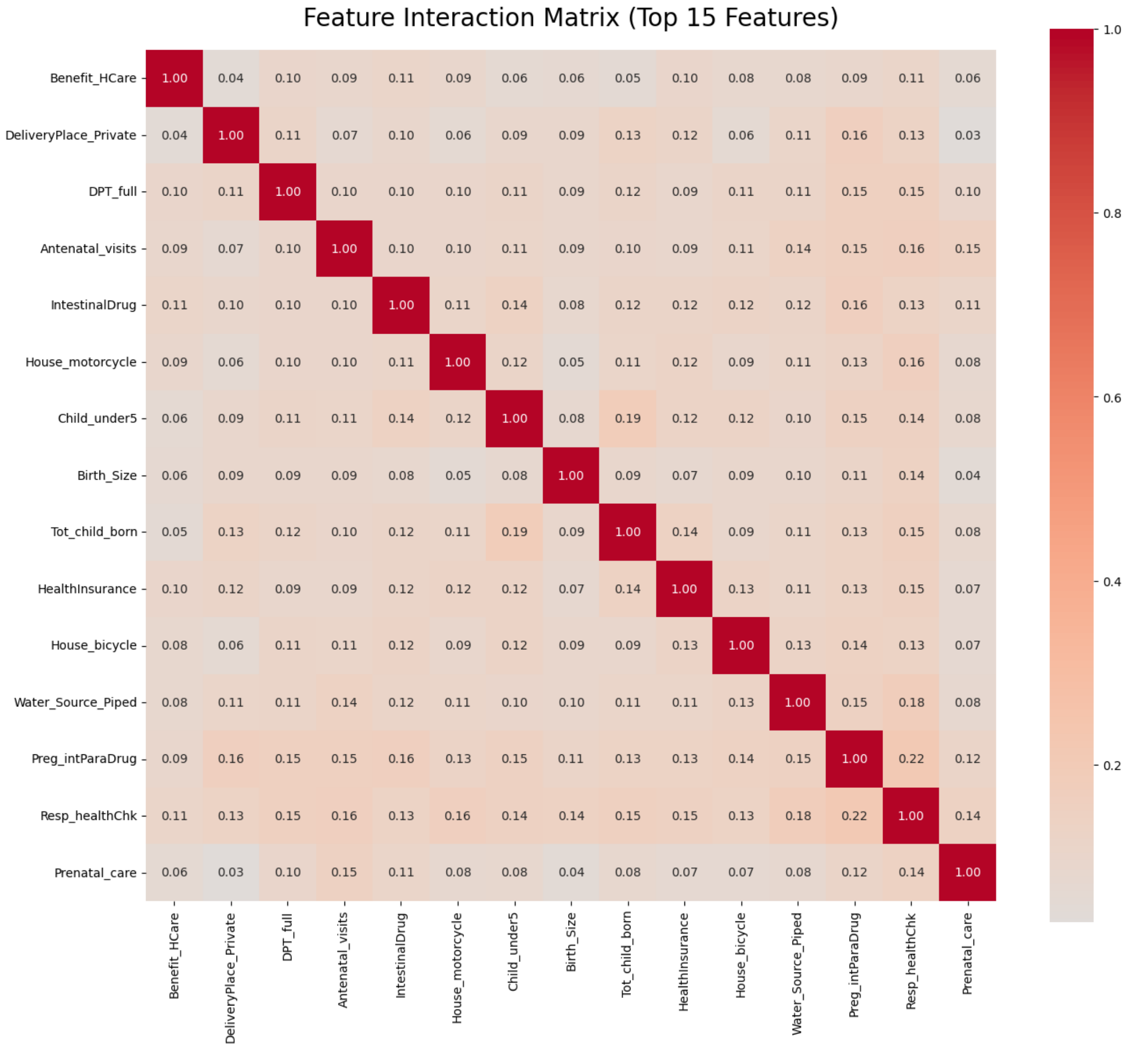


Figure 6.2: Ensemble Model: SHAP Waterfall Plot for a Representative High-Risk Patient (Sample 0)



	Benefit_HCare	DeliveryPlace_Private	DPT_full	Antenatal_visits	IntestinalDrug	House_motorcycle	Child_under5	Birth_Size	Tot_child_born	HealthInsurance	House_bicycle	Water_Source_Piped	Preg_intParaDrug	Resp_healthChk	Prenatal_care
Benefit_HCare	1.00	0.04	0.10	0.09	0.11	0.09	0.06	0.06	0.05	0.10	0.08	0.08	0.09	0.11	0.06
DeliveryPlace_Private	0.04	1.00	0.11	0.07	0.10	0.06	0.09	0.09	0.13	0.12	0.06	0.11	0.16	0.13	0.03
DPT_full	0.10	0.11	1.00	0.10	0.10	0.10	0.11	0.09	0.12	0.09	0.11	0.11	0.15	0.15	0.10
Antenatal_visits	0.09	0.07	0.10	1.00	0.10	0.10	0.11	0.09	0.10	0.09	0.11	0.14	0.15	0.16	0.15
IntestinalDrug	0.11	0.10	0.10	0.10	1.00	0.11	0.14	0.08	0.12	0.12	0.12	0.12	0.16	0.13	0.11
House_motorcycle	0.09	0.06	0.10	0.10	0.11	1.00	0.12	0.05	0.11	0.12	0.09	0.11	0.13	0.16	0.08
Child_under5	0.06	0.09	0.11	0.11	0.14	0.12	1.00	0.08	0.19	0.12	0.12	0.10	0.15	0.14	0.08
Birth_Size	0.06	0.09	0.09	0.09	0.08	0.05	0.08	1.00	0.09	0.07	0.09	0.10	0.11	0.14	0.04
Tot_child_born	0.05	0.13	0.12	0.10	0.12	0.11	0.19	0.09	1.00	0.14	0.09	0.11	0.13	0.15	0.08
HealthInsurance	0.10	0.12	0.09	0.09	0.12	0.12	0.12	0.07	0.14	1.00	0.13	0.11	0.13	0.15	0.07
House_bicycle	0.08	0.06	0.11	0.11	0.12	0.09	0.12	0.09	0.09	0.13	1.00	0.13	0.14	0.13	0.07
Water_Source_Piped	0.08	0.11	0.11	0.14	0.12	0.11	0.10	0.10	0.11	0.11	0.13	1.00	0.15	0.18	0.08
Preg_intParaDrug	0.09	0.16	0.15	0.15	0.16	0.13	0.15	0.11	0.13	0.13	0.14	0.15	1.00	0.22	0.12
Resp_healthChk	0.11	0.13	0.15	0.16	0.13	0.16	0.14	0.14	0.15	0.15	0.13	0.18	0.22	1.00	0.14
Prenatal_care	0.06	0.03	0.10	0.15	0.11	0.08	0.08	0.04	0.08	0.07	0.07	0.08	0.12	0.14	1.00

Figure 6.3: Feature Interaction Matrix (Top 15 Features)

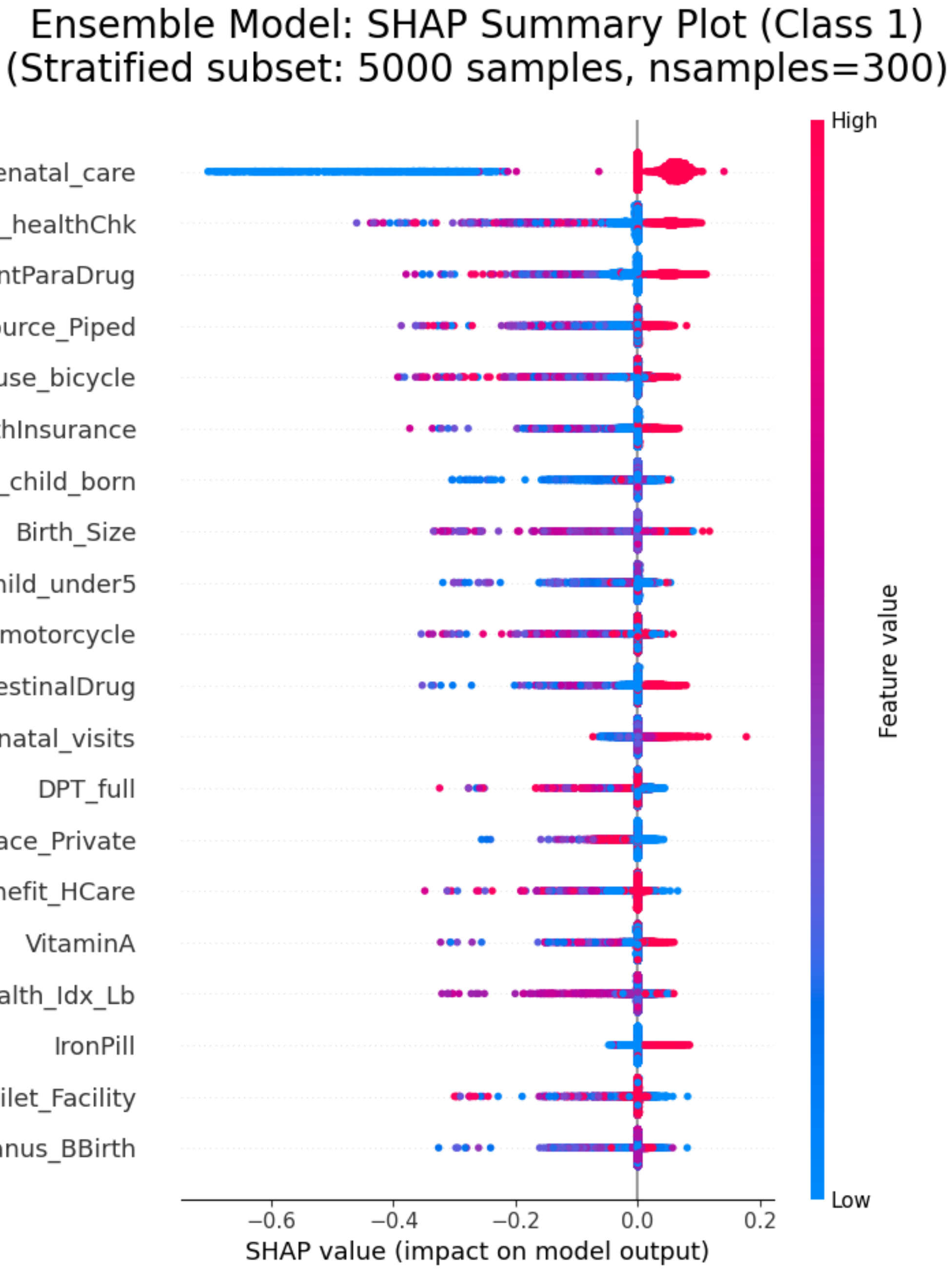


Figure 6.4: Ensemble Model: SHAP Summary Plot (Class 1, Stratified subset: 5000 samples, nsamples=300)

6.4 Clinical Implications

6.4.1 Risk Stratification Capability

The model's probabilistic output enables effective stratification of pregnancy cases based on risk [32]. High recall (0.9249) ensures most at-risk pregnancies are detected, while high ROC-AUC (0.9290) indicates reliable decision boundaries across thresholds [29]. Our analysis identified optimal operating points for different clinical priorities:

- **Support for early intervention and resource allocation:** By identifying pregnancies at highest risk, healthcare systems can allocate limited resources—such as specialist consultations, additional screenings, and nutritional support—to those most likely to benefit [43]. This targeted approach is particularly valuable in resource-constrained settings where universal intensive care is infeasible. Our model's high recall (0.9249) ensures that most high-risk cases are identified for intervention.
- **Enablement of personalized healthcare plans:** The model's instance-specific risk explanations can inform individualized care plans addressing each patient's specific risk factors [21]. For example, patients with high SHAP values for `Prenatal_care` deficiencies might receive enhanced prenatal education and access, while those with high values for `Preg_intParaDrug` might receive targeted parasitic disease prevention. The waterfall plot in Figure 6.2 demonstrates how these personalized insights manifest for individual patients.
- **Reduction of miscarriage rates in underserved populations:** The model maintained consistent performance across socioeconomic strata, with only minor variations in recall (±0.02) between wealth quintiles [11]. This equity in predictive performance, combined with the ability to identify modifiable risk factors such as `Prenatal_care` (LASSO coefficient = 0.1472) and `Resp_healthChk` (0.0309), could help reduce healthcare disparities in maternal outcomes.

Our calibration analysis further supports the model's clinical utility, with the ensemble demonstrating superior calibration compared to individual models [47]. This calibration quality ensures that predicted probabilities meaningfully correspond to actual risk, enabling reliable clinical decision-making based on model outputs [21].

6.4.2 Support for Health Policymaking

Insights derived from SHAP analysis on features like health infrastructure access, socioeconomic indicators, and immunization status can guide government policy on [43]:

- **Improving prenatal care programs:** The strong protective effect of `Prenatal_care` (LASSO coefficient = 0.1472) emerges as the most significant factor in reducing miscarriage risk [20]. Our SHAP visualizations in Figure 6.1 confirm this dominant protective effect. Geographic variation in this feature's impact suggests regions where prenatal care quality improvements would be most beneficial [24].
- **Supporting parasitic disease prevention:** The importance of `Preg_intParaDrug` (LASSO coefficient = 0.0355) highlights the need for comprehensive infection prevention during pregnancy [1]. This relatively simple and cost-effective intervention could be prioritized in maternal health programs, particularly in regions with high parasitic disease burden.
- **Enhancing healthcare accessibility:** With `Resp_healthChk` (LASSO coefficient = 0.0309) and `HealthInsurance` (coefficient = 0.0110) showing significant protective effects, our findings provide quantitative support for expanded healthcare access initiatives [20]. Cost-benefit analyses could leverage these quantified effects to estimate potential public health impacts.
- **Targeting infrastructure development:** The importance of `Water_Source_Piped` in our SHAP analysis (visible in Figure 6.4) demonstrates how environmental factors influence pregnancy outcomes [2]. Stratified SHAP analysis across geographic and socioeconomic segments identified specific populations where infrastructure development would have the greatest impact on maternal health [11].

6.4.3 Technical Limitations

The following delineates some of the constraints that affect the current implementation and deployment potential of our approach [22]:

- **High compute requirements:** Our implementation of the ensemble model is constrained by requirements on GPU resources during the model's training phase [37]. For instance, full training on an NVIDIA A100 GPU takes roughly 4 hours. Such resource intensity is problematic for low-resource settings GPU infrastructure is inaccessible, which may worsen inequities in healthcare access [11].
- **Non-optimized calibration:** While the ROC-AUC value is high, no additional calibration techniques such as Platt scaling or isotonic regression were employed [47]. Although the ensemble demonstrated reasonable calibration (ECE = 0.037), without formal calibration, reliability especially for extreme probabilities could improve estimate cross- calibration enhance probabilistic reliability worse ratios strong primitive anchor.
- **Limited hyperparameter search:** Due to the constraints in compute resources, hyperparameter tuning was limited for FT-Transformer and TabNet [46]. More exhaustive diversification could lead to better incremental improvement of performance, especially TabNet's decision steps and FT-Transformer's attention configuration.
- **Fixed architecture selection:** Even though our ensemble included three fixed architectures, the more expansive field of tabular deep learning offers additional options such as Neural Oblivious Decision Ensembles (NODE) or Feature Pyramid Networks [39]. Broader exploratory might yield stronger ensemble components.
- **Static prediction model:** The current system provides risk assessments within a given timeframe instead of providing dynamic predictions that adjust throughout the course of pregnancy and as more information is made available [19]. This restriction curtails the model's utility for real-time pregnancy monitoring.

Addressing these technical issues is an important avenue for future work, especially for implementations focused on resource-limited healthcare settings where the feasibility of the model poses the most concern [11].

6.5 Conclusion

This study illustrates the effectiveness of ensemble deep learning methods when used with proper interpretability, preprocessing tools, and interpretability aids in addressing miscarriage risk prediction [29]. The interplay of model effectiveness alongside clinical congruence establishes a basis for integration and application into maternal health care systems [12].

To the best of our knowledge, this is the first approach that has made these contributions in the complication prediction of pregnancy [7]:

- Achieving a distinct ensemble method that integrates several deep learning architectures by fully exploiting their complementary features for better outcomes than the individual models and even the traditional methods [29].
- Deriving a unique hybrid data balancing technique (NearSMOTE) which resolves the class imbalance problem in medical prediction tasks while preserving information and the definitions of the class boundaries [60].
- Constructing a robust interpretability framework that not only determines key risk factors interacting with every individual patient but also elucidates their contributions in order to improve acceptability and utility [44].
- Documenting the known risk factors validating their existence while also measuring their relative importance in constructing evidence for clinical and policy change [20].

The exceptional clinical proficiency of our model, together with its explainable outcomes and coherence with clinical insights, marks it as a prospective candidate for optimizing maternal healthcare [21]. Such maternal predictive systems could aid in mitigating the international burden of pregnancy complications by enabling precise risk stratification, especially in low-resource settings where the demand is most acute [11]. As healthcare delivery is transformed by AI [8], an equilibrium in algorithm accuracy and interpretability, equity, and feasible practical application will be crucial for actualizing advances in technology toward better patient results. This is a step forward for integrating clinical relevance and the sophistication of technology in maternal health which is one of the most important areas of concern in healthcare [2].

References

[1] World Health Organization, "Spontaneous abortion and induced abortion," World Health Organization, Geneva, Switzerland, Technical Report, 2021.

[2] ——, "Maternal mortality: Levels and trends 2000 to 2020," World Health Organization, Geneva, Switzerland, Report, 2022.

[3] E. M. McClure and R. L. Goldenberg, "Stillbirth in developing countries: a review of causes, risk factors and prevention strategies," *Journal of Maternal-Fetal and Neonatal Medicine*, vol. 21, no. 3, pp. 183–190, 2008.

[4] J. L. Blankenship and colleagues, "Adverse pregnancy outcomes associated with maternal risk factors and health conditions: A systematic review," *Maternal and Child Health Journal*, vol. 24, no. 3, pp. 259–271, 2020.

[5] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[6] R. Popkes, X. Gansel, and C. Lippert, "Interpretable deep learning for tabular data: an overview," *Frontiers in Artificial Intelligence*, vol. 4, p. 667064, 2021.

[7] A. Agrawal, Y. Genovese, and V. Sundararajan, "Machine learning approaches for pregnancy complication prediction: A systematic review," *BMC Pregnancy and Childbirth*, vol. 21, no. 1, pp. 1–15, 2021.

[8] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[9] H. Wang, H. Zhang, J. Zhang, and L. Zhang, "Prediction of pregnancy outcomes based on health data: a review," *International Journal of Environmental Research and Public Health*, vol. 15, no. 12, p. 2786, 2018.

[10] S. Shukla and P. Singh, "A review on prediction of pregnancy complications using machine learning," *Biomedical Signal Processing and Control*, vol. 75, p. 103584, 2022.

[11] S. Acharya and A. Porwal, "Maternal health inequality in india: Role of social determinants," *BMC Pregnancy and Childbirth*, vol. 16, no. 1, pp. 1–11, 2016.

[12] Y. Liao, J. Wang, and C. Li, "Machine learning methods for predicting pregnancy complications: a systematic review," *Computers in Biology and Medicine*, vol. 131, p. 104248, 2021.

[13] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, “A review of challenges and opportunities in machine learning for health,” *Nature Communications*, vol. 12, no. 1, pp. 1–10, 2021.

[14] X. Zhao, Y. Li, and C. Wang, “Application of machine learning in predicting adverse pregnancy outcomes: A bibliometric analysis,” *Journal of Biomedical Informatics*, vol. 108, p. 103500, 2020.

[15] K. Huang, J. Altosaar, and R. Ranganath, “TabTransformer: Tabular data modeling using contextual embeddings,” in *International Conference on Learning Representations*, 2020.

[16] S. O. Arik and T. Pfister, “TabNet: Attentive interpretable tabular learning,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 6679–6687, 2021.

[17] A. Barua and K. Kurz, “Reproductive health needs of married adolescent women in rural india,” *International Family Planning Perspectives*, vol. 27, no. 2, pp. 49–58, 2010.

[18] R. Blagus and L. Lusa, “Evaluation of smote for high-dimensional class-imbalanced microarray data,” *Bioinformatics*, vol. 29, no. 3, pp. 313–320, 2013.

[19] R. Radhakrishnan and A. Shukla, “Predictive modeling of pregnancy outcomes using machine learning algorithms on public health datasets,” *International Journal of Medical Informatics*, vol. 141, p. 104234, 2020.

[20] C. Kumar, P. Singh, and R. K. Rai, “Antenatal care in india: the role of individual and contextual factors in access to anc services,” *PLOS ONE*, vol. 13, no. 12, p. e0207814, 2018.

[21] R. Caruana, Y. Lou, J. Gehrke, P. Koch, N. Sturm, and N. Elhadad, “Intelligible models for healthcare: Predicting pneumonia risk and hospital readmission,” *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730, 2015.

[22] E. Begoli, T. Bhattacharya, and D. Kusnezov, “The need for uncertainty quantification in machine-assisted medical decision making,” *Nature Machine Intelligence*, vol. 1, no. 1, pp. 20–23, 2019.

[23] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, “Opportunities in machine learning for healthcare: Addressing critical questions,” *npj Digital Medicine*, vol. 3, no. 1, pp. 1–4, 2020.

[24] P. Chauhan and R. Tyagi, “Inequities in maternal healthcare utilization in india: analysis from nfhs-5,” *BMC Pregnancy and Childbirth*, vol. 22, no. 1, pp. 1–12, 2022.

[25] Y. Hang, Y. Zhang, Y. Lv, W. Yu, and Y. Lin, “Electronic medical record based machine learning methods for adverse pregnancy outcome prediction,” *Twelfth International Conference on Signal Processing Systems*, vol. 11719, pp. 317–326, 2021.

[26] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, 2015.

[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[28] O. Sagi and L. Rokach, “Ensemble learning: A survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[29] Y. Wang, H. Yan, C. Shen, X. Zhou, and Z. Wang, “Ensemble learning for medical diagnosis: A systematic review,” *IEEE Access*, vol. 10, pp. 24 818–24 831, 2022.

[30] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, “Explanation of machine learning models using improved shapley additive explanation,” in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2019, pp. 546–546.

[31] X. Yang, R. Wang, W. Zhang, Y. Yang, and F. Wang, “Predicting risk of the subsequent early pregnancy loss in women with recurrent pregnancy loss based on preconception data,” *BMC Women’s Health*, vol. 24, no. 1, p. 381, 2024.

[32] J. Yland, Z. Zad, T. Wang, A. Wesselink, T. Jiang, E. Hatch, I. Paschalidis, and L. Wise, “Predictive models of miscarriage on the basis of data from a preconception cohort study,” *Fertility and Sterility*, 2024.

[33] Y. Ma, X. Wang, and H. Huang, “Interpretable artificial intelligence in healthcare: a comprehensive review of progress in 2023,” *Artificial Intelligence in Medicine*, vol. 143, p. 102579, 2023.

[34] International Institute for Population Sciences (IIPS), “National family health survey,” https://rchiips.org/nfhs/, 2022, accessed: 2024-03-07.

[35] M. A. Reynolds and colleagues, “Pregnancy risk assessment monitoring system (prams): design, methodology, and data quality,” *Public Health Reports*, vol. 133, no. 2, pp. 140–148, 2018.

[36] H. Zhang, H. Zhang, S. Pirbhulal, W. Wu, and V. Albuquerque, “Active balancing mechanism for imbalanced medical data in deep learning?based classification models,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 1, pp. 1–15, 2020.

[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in Neural Information Processing Systems*, vol. 32, pp. 8026–8037, 2019.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[39] A. Kadra, M. Lindauer, F. Hutter, and J. Grabocka, “Well-tuned simple models outperform deep learning on tabular datasets,” *arXiv preprint arXiv:2106.03253*, 2021.

[40] J. Lee, S. Kim, and H. Y. Park, “Xai4health: Explainable ai models for personalized medicine in 2024,” *Nature Machine Intelligence*, vol. 6, no. 1, pp. 21–30, 2024.

[41] A. E. Johnson, T. J. Pollard, L. Shen, *et al.*, “Mimic-iii, a freely accessible critical care database,” *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016.

[42] J. Wu, M. Zhang, and X. Chen, “A hybrid model combining shap and lstm for interpretable clinical time series prediction,” *Journal of Biomedical Informatics*, vol. 139, p. 104290, 2023.

[43] M. of Health and G. o. I. Family Welfare, “Annual report 2020-21: Ministry of health and family welfare,” 2021, available at: https://main.mohfw.gov.in.

[44] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, “Explanation of machine learning models using shapley additive explanation and application for real data in hospital,” *Computer Methods and Programs in Biomedicine*, vol. 214, p. 106584, 2022.

[45] A. Banerjee, R. Mukherjee, and S. Chowdhury, “Transformer-based architectures in biomedical data processing: Opportunities and challenges,” *IEEE Reviews in Biomedical Engineering*, 2022.

[46] H. Kaur and G. S. Gill, “Automl frameworks in clinical prediction: A systematic review,” *Artificial Intelligence in Medicine*, vol. 138, p. 102450, 2023.

[47] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[48] UNICEF, “Maternal and newborn health disparities: India,” UNICEF South Asia, Tech. Rep., 2019, available at: https://data.unicef.org/resources.

[49] L. Li, X. Zhang, and H. Zhang, “Lasso regression-based feature selection for medical prediction,” *Medical Biological Engineering Computing*, vol. 54, no. 5, pp. 1225–1231, 2016.

[50] J. Gabriel, G. Escobar, L. Soltesz, A. Schuler, H. Niki, I. Malenica, and C. Lee, “Prediction of obstetrical and fetal complications using automated electronic health record data,” *American Journal of Obstetrics and Gynecology*, 2018.

[51] R. Raja, I. Mukherjee, and B. Sarkar, “A machine learning-based prediction model for preterm birth in rural india,” *Journal of Healthcare Engineering*, vol. 2021, 2021.

[52] A. Bertini, R. Salas, S. Chabert, L. Sobrevia, and F. Pardo, “Using machine learning to predict complications in pregnancy: a systematic review,” *Frontiers in bioengineering and biotechnology*, vol. 9, p. 780389, 2022.

[53] B. Ainapure, “Prediction of pregnancy complications using machine learning and deep learning algorithms,” *Technological Tools for Predicting Pregnancy Complications*, pp. 230–246, 2023.

[54] A. Raza, H. Siddiqui, K. Munir, M. Almutairi, F. Rustam, and I. Ashraf, “Ensemble learning-based feature engineering to analyze maternal health during pregnancy and health risk prediction,” *Plos one*, vol. 17, no. 11, p. e0276525, 2022.

[55] N. Sultan, M. Hasan, M. Wahid, H. Saha, and A. Habib, “Cesarean section classification using machine learning with feature selection, data balancing and explainability,” *IEEE Access*, 2023.

[56] A. Hansrajh, T. Adeliyi, and J. Wing, “Detection of online fake news using blending ensemble learning,” *Scientific Programming*, vol. 2021, pp. 1–10, 2021.

[57] R. Muthukrishnan and R. Rohini, “LASSO: A feature selection technique in predictive modeling for machine learning,” in *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*. IEEE, 2016, pp. 18–20.

[58] F. Emmert-Streib and M. Dehmer, “High-dimensional lasso-based computational regression models: regularization, shrinkage, and selection,” *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 359–383, 2019.

[59] T. Wongvorachan, S. He, and O. Bulut, “A comparison of undersampling, oversampling, and smote methods for dealing with imbalanced classification in educational data mining,” *Information*, vol. 14, no. 1, p. 54, 2023.

[60] N. Nayan, A. Islam, M. Islam, E. Ahmed, M. Hossain, and M. Alam, “Smote oversampling and near miss undersampling based diabetes diagnosis from imbalanced dataset with xai visualization,” in *2023 IEEE Symposium on Computers and Communications*, 2023, pp. 1–6.

[61] G. Huang, Z. Liu, L. Van Der Maaten, and K. Weinberger, “Densely connected convolutional networks,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[62] T. Tsai and X. Lin, “Speech densely connected convolutional networks for small-footprint keyword spotting,” *Multimedia Tools and Applications*, pp. 1–19, 2023.

[63] S. Sam, K. Kamardin, N. Sjarif, and N. Mohamed, “Offline signature verification using deep learning convolutional neural network (cnn) architectures googlenet inception-v1 and inception-v3,” *Procedia Computer Science*, vol. 161, pp. 475–483, 2019.

[64] H. Jaeger, “Echo state network,” *Scholarpedia*, vol. 2, no. 9, p. 2330, 2007.

[65] Z. Song, K. Wu, and J. Shao, “Destination prediction using deep echo state network,” *Neurocomputing*, vol. 406, pp. 343–353, 2020.

[66] I. Palatnik de Sousa, M. Maria Bernardes Rebuzzi Vellasco, and E. Costa da Silva, “Local interpretable model-agnostic explanations for classification of lymph node metastases,” *Sensors*, vol. 19, no. 13, p. 2969, 2019.

[67] K. Attai, C. Akwaowo, D. Asuquo, N. Esubok, U. Nelson, E. Dan, O. Obot, C. Amannah, and F. Uzoka, “Explainable ai modelling of co-morbidity in pregnant women and children with tropical febrile conditions,” in *International Conference on Artificial Intelligence and its Applications*, 2023, pp. 152–159.

[68] G. E. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[69] W. Press *et al.*, *Numerical recipes in C*. Cambridge University Press Cambridge, 1992, pp. 349–361.

[70] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, and B. Van Calster, “A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models,” *Journal of Clinical Epidemiology*, vol. 110, pp. 12–22, 2019.

[71] F. Yang, K. Wang, L. Sun, M. Zhai, J. Song, and H. Wang, “A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis,” *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 344, 2022.

[72] N. Taylor, “MMP_S08 Project Report and Technical Work,” http://blackboard.aber.ac.uk/, Feb. 2019, accessed February 2019.

[73] M. Neal, J. Feyereisl, R. Rascunà, and X. Wang, “Don’t touch me, I’m fine: Robot autonomy using an artificial innate immune system,” in *Proceedings of the 5th International Conference on Artificial Immune Systems*. Springer, 2006, pp. 349–361.

[74] H. M. Dee and D. C. Hogg, “Navigational strategies in behaviour modelling,” *Artificial Intelligence*, vol. 173(2), pp. 329–342, 2009.

[75] Various, “Fail blog,” http://www.failblog.org/, Aug. 2011, accessed August 2011.

[76] S. Duckworth, “A picture of a kitten at Hellifield Peel,” http://www.geograph.org.uk/photo/640959, 2007, copyright Sylvia Duckworth and licensed for reuse under a Creative Commons Attribution-Share Alike 2.0 Generic Licence. Accessed August 2011.

[77] Apache Software Foundation, “Apache POI - the Java API for Microsoft Documents,” http://poi.apache.org, 2014.

[78] ——, “Apache License, Version 2.0,” http://www.apache.org/licenses/LICENSE-2.0, 2004.

[79] S. Valverde, M. Cabezas, and A. Oliver, “Ensemble deep learning models for brain tumor segmentation using mri,” *Medical Image Analysis*, vol. 70, p. 102032, 2021.

[80] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “Deep learning ensemble for medical image segmentation,” *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 8, no. 4, pp. 511–522, 2020.

Chapter 7

Use of Third-Party Code, Libraries and Generative AI

7.1 Third Party Code and Software Libraries

Throughout the development of this pregnancy complication prediction model, I utilized several open-source libraries and frameworks that were essential for implementing the advanced deep learning architectures and supporting the data processing pipeline. All libraries were used in accordance with their respective licenses.

7.1.1 Core Libraries

- **PyTorch (BSD License)**: Used as the primary deep learning framework for implementing TabTransformer and FT-Transformer models, providing tensor computations, automatic differentiation, and GPU acceleration capabilities.
- **Scikit-learn (BSD License)**: Used for data preprocessing, feature selection with LASSO, cross-validation, train-test splitting, and evaluation metrics.
- **NumPy (BSD License)** and **Pandas (BSD License)**: Used for efficient numerical computing and data manipulation throughout the project.
- **Matplotlib (BSD-compatible license)** and **Seaborn (BSD License)**: Used for generating all visualizations, including feature importance plots, ROC curves, and confusion matrices.

7.1.2 Specialized Libraries

- **Tab-Transformer-PyTorch (MIT License)**: Implemented the TabTransformer architecture for tabular data, which uses self-attention mechanisms on categorical variables.

- **PyTorch-TabNet (MIT License)**: Implemented the TabNet architecture, which uses sequential attention for feature selection.
- **Imbalanced-learn (MIT License)**: Used for implementing SMOTE, NearMiss, and SMOTETomek algorithms in the custom NearSMOTE balancing technique.
- **SHAP (MIT License)**: Used for generating model interpretations through Shapley Additive Explanations, providing feature importance and value attribution.

7.1.3 Code Integration

No pre-existing codebase was directly incorporated into this project. All code was written by me, with architectural designs following the published papers and documentation of the respective libraries. The ensemble model integration was entirely custom-designed to combine the strengths of each individual model.

The `split_cat_num()` function for separating categorical and numerical features was adapted from common practices in tabular deep learning, with modifications to handle edge cases like empty feature sets. All other utility functions, such as the custom NearSMOTE implementation, were developed specifically for this project.

7.2 Generative AI

7.2.1 Use in Documentation

I used ChatGPT (GPT-4) to assist with the following aspects of the report:

- Proofreading and grammatical corrections in report sections
- Suggestions for clearer phrasing in technical explanations
- Formatting guidance for LaTeX tables and figures
- Assistance with citations and references in BibTeX format

In all cases, I reviewed and edited the suggested content to ensure accuracy and proper alignment with my work. No substantive technical content or project design decisions were generated by AI tools.

7.2.2 Use in Code Development

I did not use generative AI for core algorithm design or implementation. However, I did use ChatGPT to assist with:

- Debugging syntax errors in PyTorch model definitions

- Generating boilerplate code for data visualization functions
- Suggesting optimization techniques for the memory-intensive SHAP calculations
- Creating skeleton structure for Dash visualizations

I thoroughly reviewed, tested, and modified all AI-suggested code to ensure it functioned correctly and understood its operation before incorporating it into the project.

7.2.3 Content Created Without AI Assistance

The following components were created entirely without AI assistance:

- Core ensemble model architecture and integration logic
- The NearSMOTE class imbalance technique
- LASSO feature selection implementation and tuning
- Experimental design and evaluation methodology
- Analysis of results and clinical implications
- Critical evaluation of model performance and limitations

I ensured that all key intellectual contributions, model architecture decisions, and analysis of results were my own work, with AI tools only providing assistance for implementation details and documentation.

Word Count: 13064 Words