Integrating Healthcare, Financial, and Insurance Data for Predictive Modeling and Pattern Mining of Diabetes Medication Usage

1st Reem Hussin Mostafa
Nile University
Egypt, Cairo
R.Hussin2241@nu.edu.eg

2th Mahmoud Essa
Nile University
Egypt, Cairo
M.Ahmed2219@nu.edu.eg

3rd Mohanned Mahmoud
Nile University
Egypt, Cairo
M.Mahmoud2207@nu.edu.eg

4th Abdelrahman Ahmed
Nile university
Egypt, Cairo
A.Shreif2299@nu.edu.eg

Abstract— In this study, we present an integrated data mining approach to predict and analyze diabetes medication usage by combining patient health records, insurance information, and financial background data. The dataset was constructed by linking hospital treatment data, insurance provider details, and banking attributes using shared identifiers such as PatientID and AccountID. To predict whether a patient requires diabetes medication, we applied a Random Forest classifier, achieving an accuracy of 83%, with key contributing features including age, A1C test results, glucose levels, and diagnosis codes. In parallel, a Decision Tree model was implemented to offer an interpretable view of the decision process, providing an accuracy of 80% while visually highlighting key decision splits such as medication change, readmission status, and medication counts. In addition, association rule mining using the Apriori algorithm was employed to uncover frequent patterns and risk factors associated with diabetes treatment, revealing high-confidence relationships involving emergency visits, medical specialties, and socioeconomic indicators. The combination of supervised classification and unsupervised pattern discovery provides both predictive power and interpretability. This work demonstrates the value of multi-domain data integration in healthcare analytics, offering actionable insights for hospitals, insurance providers, and policymakers in identifying high-risk patients and improving treatment planning.

Keywords
Apriori, Data Mining, Machine Learning, Classification, diabetes, association rules, frequent patterns, Decision Tree

I. INTRODUCTION

Diabetes mellitus has emerged as a significant global health concern, with its prevalence escalating at an alarming rate. According to the International Diabetes Federation (IDF), approximately 589 million adults aged 20–79 years were living with diabetes in 2024, a figure projected to rise to 853 million by 2050 [1]. This surge is predominantly attributed to type 2 diabetes, which accounts for over 90% of all diabetes cases and is closely linked to factors such as urbanization, sedentary lifestyles, and dietary habits.

The burden of diabetes extends beyond health implications, imposing substantial economic strains on individuals and healthcare systems. In 2024, diabetes-related health expenditures reached an estimated USD 1.015 trillion globally, marking a 338% increase over the past 17 years [1]. In regions like Punjab, India, studies have highlighted the catastrophic health expenditures faced by households due to non-communicable diseases, with diabetes patients experiencing the highest incidence [2].

Early detection and effective management of diabetes are crucial in mitigating its complications, which include cardiovascular diseases, kidney failure, and neuropathy [3]. Traditional clinical assessments, while essential, often fall short in capturing the multifaceted nature of diabetes risk factors, especially when considering socio-economic and behavioral determinants.

In recent years, data mining and machine learning techniques have been increasingly employed to enhance predictive modeling in healthcare. These approaches facilitate the analysis of large and complex datasets, uncovering patterns and associations that might remain undetected through conventional methods [4]. Specifically, in the context of diabetes, predictive models have been developed to estimate the probability of disease onset, progression, and the necessity for medication, thereby supporting clinical decision-making and personalized treatment plans.

Building upon this foundation, our study aims to integrate healthcare, financial, and insurance data to develop predictive models and extract meaningful patterns related to diabetes medication usage. By combining diverse data sources, we seek to capture a holistic view of the factors influencing diabetes management, encompassing medical history, socio-economic status, and insurance coverage. Through the application of Random Forest classification and Apriori association rule mining, we endeavor to identify key predictors and associations that can inform targeted interventions and policy decisions.

Building upon this foundation, our study aims to integrate healthcare, financial, and insurance data to develop predictive models and extract meaningful patterns related to diabetes medication usage. By combining diverse data sources, we

seek to capture a holistic view of the factors influencing diabetes management, encompassing medical history, socio-economic status, and insurance coverage. Through the application of Random Forest classification, Decision Tree analysis, and Apriori association rule mining, we endeavor to identify key predictors and associations that can inform targeted interventions and policy decisions.

II.RELATED WORK

- *Machine Learning in Diabetes Prediction*

The application of machine learning (ML) techniques in diabetes prediction has been extensively explored in recent years. Traditional models, such as decision trees, logistic regression, and naïve Bayes classifiers, have been employed to identify individuals at risk of developing type 2 diabetes mellitus (T2DM) based on clinical and demographic factors [1]. More advanced models, including support vector machines (SVM), random forests (RF), and deep learning architectures, have demonstrated improved predictive performance [2][13].

A comprehensive review by Kiran et al. highlighted the evolution of ML and artificial intelligence (AI) applications in T2DM prediction over the past three decades, emphasizing the shift towards more complex models and the integration of diverse data sources [3]. These models have been instrumental in developing clinical decision support systems (CDSS) that aid healthcare providers in early diagnosis and personalized treatment planning [4].

Recent studies have also focused on the interpretability of ML models. For instance, Alam et al. developed an automatic diabetes prediction system using various ML techniques and employed explainable AI approaches like LIME and SHAP to understand model predictions [5]. Such interpretability is crucial for clinical adoption and trust in AI-driven healthcare solutions.

- *Association Rule Mining in Diabetes Research*

Association rule mining (ARM) has been utilized to uncover hidden patterns and relationships among various risk factors associated with diabetes. Ramezankhani et al. applied ARM to the Tehran Lipid and Glucose Study database, identifying combinations of anthropometric and lifestyle factors that significantly increased the risk of T2DM [6]. Similarly, Simon et al. introduced Survival Association Rule (SAR) mining, an extension of traditional ARM, to assess T2DM risk by accounting for time-to-event data and confounding variables [7].

Further research by Senthamarai et al. employed ARM based on clustering to estimate diabetes risk, categorizing patients into high, medium, and low-risk groups [8]. Additionally, Narindrarangkura et al. utilized ARM on real-world data to uncover links between race, glycemic control, lipid profiles, and suicide attempts in individuals with diabetes, highlighting the multifaceted nature of diabetes-related complications [9].

- *Integration of Multidomain Data for Enhanced Prediction*

Integrating data from multiple domains—such as healthcare, financial, and insurance records—has been shown to improve the accuracy and applicability of predictive models for diabetes. Combining clinical data with socioeconomic and behavioral information allows for a more comprehensive understanding of the factors influencing disease onset and progression [10]. For example, studies have shown that incorporating insurance claims and financial data can enhance the prediction of medication adherence and disease management outcomes [11].

However, challenges remain in harmonizing data from disparate sources, addressing privacy concerns, and ensuring data quality. Despite these obstacles, the integration of multidomain data holds promise for developing more robust and personalized predictive models for diabetes care [12].

III.METHODOLOGY

This research followed a structured data engineering and analytics pipeline, encompassing data integration, warehousing, transformation, visualization, and advanced analysis using classification and association rule mining. The pipeline begins with the raw data ingestion from hospital, bank, and insurance domains, followed by the ETL process (Extract, Transform, Load) and the construction of a star schema for efficient querying. This process culminates in data analysis using both classification models and association rule mining to derive meaningful insights.

The end-to-end workflow of the project is summarized in Figure 1, which visually represents the entire methodology, from data ingestion through to the final analysis.

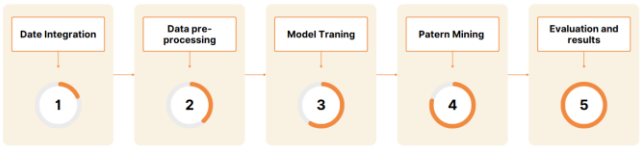


Figure 1. End-to-End Workflow of the Project

A. Data Sources

In this study, we utilized two key datasets to develop and evaluate our models:

1. European Central Bank (ECB) Data: The ECB dataset is a comprehensive collection of financial and economic data, which includes various indicators related to the European economy. The data is publicly available via the ECB Statistical Data Warehouse [17] and provides valuable insights into financial factors that can complement healthcare data in predictive modeling. This dataset will be used to incorporate financial variables in the analysis, such as economic growth, inflation rates, and interest rates, which can significantly influence health outcomes and treatment decisions.
2. UCI Diabetes Dataset: The UCI Diabetes Dataset, titled Diabetes 130-US hospitals for years 1999-2008, is sourced from the UCI Machine Learning Repository. This dataset contains clinical data from diabetes patients across 130 hospitals in the United States, covering a period from 1999 to 2008. It includes demographic, clinical, and historical medical data such as age, gender,

A1C test results, and medication history, which are critical for understanding treatment decisions for diabetes medication. This dataset is instrumental in evaluating the effectiveness of predictive models for healthcare outcomes and is used to train and test our Random Forest and Decision Tree models.

B. Data Integration and Warehousing

To systematically manage raw healthcare, insurance, and financial data, a Medallion Architecture approach was adopted using a SQL Server-based data warehouse. This architecture consists of three layers: Bronze, Silver, and Gold, each responsible for specific roles in the ETL pipeline.

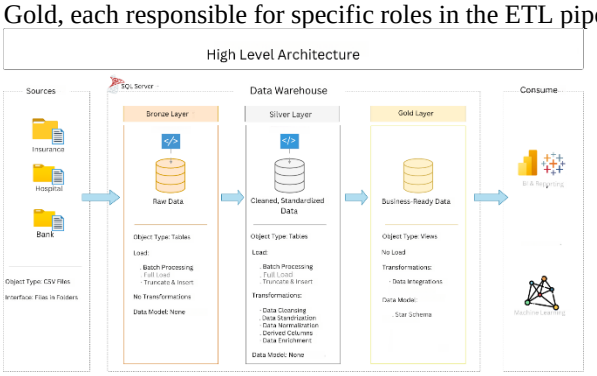


Figure 2. Medallion Architecture Overview

In this diagram we highlight the definition, objectives, and processes for each layer.

- Bronze Layer: Ingested raw CSV files from hospital, insurance, and bank systems. Data was loaded as-is using BULK INSERT into raw staging tables.
- Silver Layer: Performed data cleansing, type normalization, column derivation, and basic validations using stored procedures.
- Gold Layer: Designed for analytics consumption; this layer aggregated clean data into dimensional models (e.g., DimAccount, DimMedicalEncounterAttributes, FactHospitalEncounter) using star schema structures.

Data engineers were responsible for designing and managing these layers, ensuring proper traceability, full-load management (truncate & insert), and consistent transformation logic throughout the process.

C. Data Preparation and Feature Engineering

Once the data was loaded and structured across the Medallion architecture, the relevant entities were joined into a single integrated schema using Power BI's modeling layer. This relationship model enabled seamless joining across hospital visits, patients, insurance records, and financial profiles through shared keys like AccountID and InsuranceID.

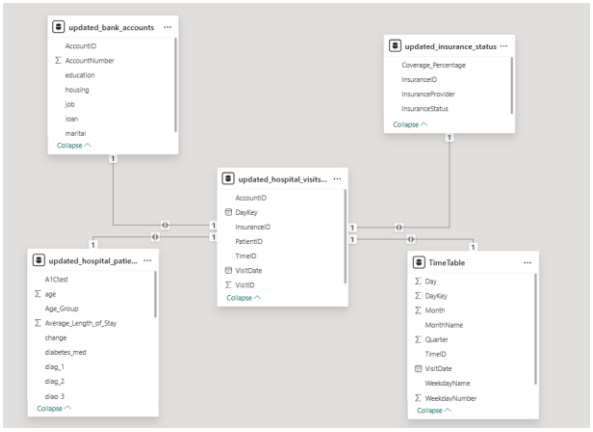


Figure 3. Semantic Model Overview

we visual represents the semantic model where:

- hospital_visits acts as the fact table
- and hospital_patients, insurance_status, bank_accounts, and TimeTable serve as dimension tables

These relationships were the foundation for both machine learning modeling and association rule mining, allowing the unified selection of features from healthcare, financial, and demographic domains.

After merging the datasets through common identifiers (AccountNumber, InsuranceID), the unified table included demographic, clinical, and insurance variables. This merged dataset was modeled within Power BI using a star schema, where hospital_visits functioned as the fact table and other tables—such as hospital_patients, insurance_status, bank_accounts, and TimeTable—served as dimensions.

The result was a flattened table that included all necessary features selected across domains. This table was exported and used as the input dataset for subsequent machine learning and pattern mining tasks. Key features included: age, A1Ctest, glucose_test, change, medical_specialty, n_medications, time_in_hospital, readmitted, education, job, loan, InsuranceProvider, and InsuranceStatus.

This process ensured that the dataset was not only clean and integrated but also rich in attributes that span healthcare, insurance, and financial dimensions—ideal for predictive and associative modeling.

D. Predictive Modeling and Pattern Mining

Following the integration and preprocessing of the healthcare, financial, and insurance datasets, two key data mining techniques were employed to derive meaningful insights and support predictive analytics: supervised classification and unsupervised association rule mining.

1. Supervised Classification

To predict whether a patient would be prescribed diabetes medication, we implemented a supervised learning model using the Random Forest Classifier. This ensemble-based algorithm is particularly suitable for handling heterogeneous tabular data containing both numerical and categorical

features. The model was trained on a refined feature set extracted from the merged dataset, comprising demographic, medical, and financial attributes. Specifically, the selected features included:
age, A1Ctest, glucose_test, change, diag_1, diag_2, diag_3, medical_specialty, n_medications, time_in_hospital, readmitted, education, job, loan, InsuranceProvider, and InsuranceStatus.

Categorical features were preprocessed using label encoding, while numerical features were left in their original form due to the robustness of tree-based models to unscaled data. The dataset was split into training and testing subsets using an 80:20 stratified sampling approach to preserve class distribution.

The model was evaluated using common classification metrics including accuracy, precision, recall, F1-score, and confusion matrix. These metrics were selected to provide a comprehensive understanding of the classifier's performance, particularly its ability to correctly identify positive prescriptions without bias toward the majority class. The results of this evaluation are detailed in Section 4.

To complement the ensemble-based Random Forest model, a Decision Tree classifier was also implemented to provide an interpretable baseline model. With a limited depth to avoid overfitting, the Decision Tree enabled direct visualization of key decision rules used to classify diabetes medication prescriptions. This transparency is particularly beneficial for understanding clinical decision pathways and communicating model logic to non-technical stakeholders. The results of the Decision Tree model, including pruned visualization and comparative performance metrics, are also reported in Section 4.

2. Unsupervised Association Rule Mining

In addition to predictive modeling, unsupervised pattern discovery was performed using the Apriori algorithm to identify strong associative relationships between patient characteristics and the prescription of diabetes medication. Each patient record was transformed into a transaction of categorical attributes, effectively enabling the construction of frequent itemsets that reveal co-occurrence patterns.

To ensure meaningful and interpretable rule generation, the mining process was constrained by minimum thresholds for support ($\geqslant 0.01$), confidence ($\geqslant 0.5$), and lift ($\geqslant 1$). Rules were extracted in the form of:

Antecedents ⇒ Consequent (diabetes_med = yes)

These association rules provided interpretable insights into factors frequently associated with medication usage. For instance, combinations of chronic conditions (e.g., circulatory diagnosis), financial indicators (e.g., no loan, specific job sectors), and test outcomes (e.g., A1Ctest = yes) were found to frequently co-occur with diabetes prescriptions. Unlike classification, this approach uncovered previously unknown correlations, complementing the supervised model by offering exploratory and explanatory perspectives on the underlying data.

II. RESULTS AND DISCUSSION

A. Predictive Insights and Model Outcomes

To assess the predictive capability of our integrated healthcare-financial dataset, we employed the Random Forest classifier to predict whether a patient would be prescribed diabetes medication. This model was chosen due to its robustness against overfitting and its ability to handle both numerical and categorical variables effectively.

The model was trained on a curated dataset that merged critical features from hospital visits (e.g., diagnosis codes, A1C tests, medication history), patient demographics (e.g., age, education, job), and insurance information (e.g., provider, status). This multidimensional approach was intended to simulate real-world predictive scenarios where both clinical and socioeconomic variables influence prescription decisions.

The confusion matrix shown in Figure 4 provides a visual overview of the model's classification results.

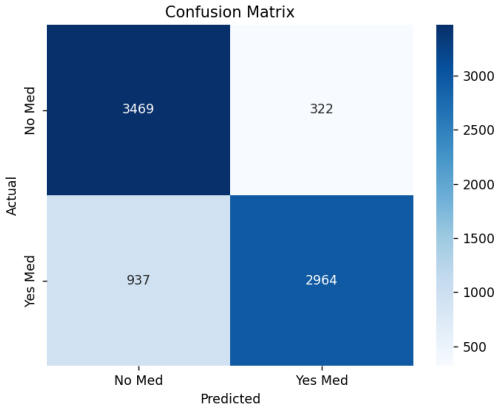


Figure 4. Confusion Matrix of Random Forest Classifier

From the matrix:

- True Positives (TP): 2,964 patients were correctly predicted to receive medication.
- True Negatives (TN): 3,469 patients were correctly identified as not requiring medication.
- False Positives (FP): 322 patients were wrongly predicted to need medication.
- False Negatives (FN): 937 patients were incorrectly predicted not to receive medication.

These results demonstrate that the model has strong overall performance, particularly in identifying non-medicated cases. However, the presence of false negatives suggests some challenges in capturing all patients who need medication—an expected limitation in imbalanced healthcare data where many clinical and behavioral factors affect treatment decisions.

To further assess the model's performance, we analyzed key evaluation metrics including precision, recall, and F1-score for each class, as displayed in Table 1:

Class Label	*Precision*	*Recall*	*F1-Score*	*Support*
No Med (0)	*0.79*	*0.92*	*0.85*	*3791*
Yes Med (1)	*0.90*	*0.76*	*0.82*	*3901*

Accuracy			*0.84*	*7692*
Macro Avg	*0.84*	*0.84*	*0.84*	*7692*
Weighted Avg	*0.85*	*0.84*	*0.84*	*7692*

Table 1. Model Evaluation Metrics

- Precision for the class "Yes Med" reached 0.90, indicating a high proportion of patients predicted to receive medication did receive it.
- Recall was 0.76 for "Yes Med", which reflects that 76% of all actual diabetes-medicated cases were successfully identified by the model.
- F1-score, the harmonic mean of precision and recall, was 0.82 for "Yes Med", and 0.85 for "No Med", showing balanced performance across both outcomes.
- Overall accuracy stood at 84%, which is notably strong for a healthcare dataset involving behavior and treatment variability.

These metrics underscore that the model is particularly strong in minimizing false positives and demonstrates respectable ability to detect medicated patients as a vital factor in proactive healthcare management.

In addition, we analyzed the importance of determining which variables contributed most significantly to the model's decisions. Figure 5 illustrates the ranked importance of features in predicting diabetes medication usage:

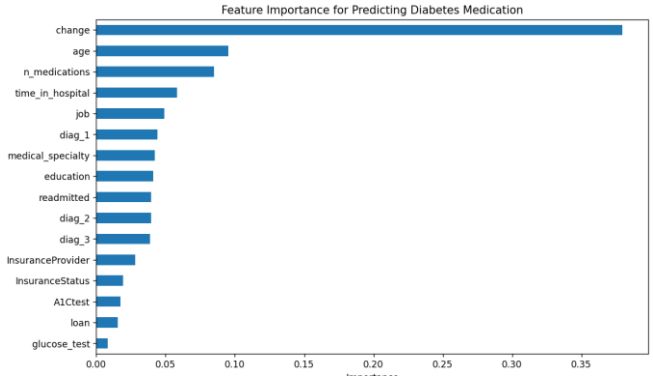


Figure 5. Feature Importance in Random Forest Model

From this, the most influential features were:

- Change in medication status: The dominant predictor, reflecting whether treatment had been modified during the encounter.
- Age and Number of Medications: Indicating the relevance of chronicity and complexity of patient care.
- Time in Hospital, Job, and Primary Diagnosis: Socio-clinical variables influencing long-term prescriptions.

This analysis highlights the strength of integrating cross-domain data: clinical, demographic, and financial variables jointly improve model interpretability and decision-making.

B. Interpretable Classification via Decision Tree

To enhance model interpretability and provide a transparent representation of the decision-making process, we also implemented a Decision Tree classifier with a constrained depth. This model offers visual insight into how feature values lead to prediction outcomes, which is especially useful for healthcare professionals seeking traceable logic in AI systems.

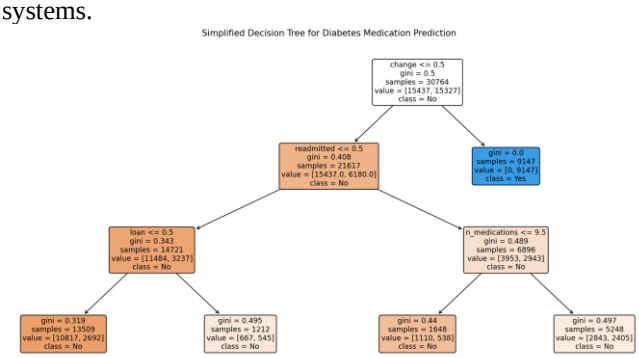


Figure 6. Trained Decision Tree Model

The trained tree (shown in Figure 6) highlights the most influential split features, such as change, n_medications, and diag_1, and terminates in leaf nodes predicting whether a patient will be prescribed diabetes medication. The tree structure provides intuitive explanations — for example, patients who experienced a change in treatment and had more than four medications been often classified as positive cases.

The performance of the Decision Tree model is summarized in Table 2. Although its accuracy (81%) is slightly lower than the Random Forest, the simplicity of the model makes it valuable for explanatory and clinical decision support purposes.

Class Label	*Precision*	*Recall*	*F1-Score*	*Support*
No Med (0)	*0.76*	*0.88*	*0.81*	*3791*
Yes Med (1)	*0.87*	*0.73*	*0.79*	*3901*
Accuracy			*0.81*	*7692*
Macro Avg	*0.82*	*0.81*	*0.80*	*7692*
Weighted Avg	*0.82*	*0.81*	*0.80*	*7692*

Table 2. Performance Summary of Decision Tree Model

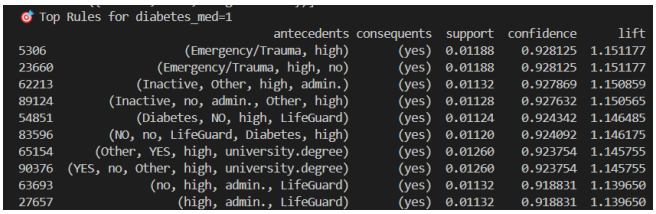
C. Insights from Association Rule Mining

To extract meaningful patterns that co-occur with the prescription of diabetes medication, we applied the Apriori algorithm on the transformed dataset. This unsupervised learning technique helps in identifying frequent itemsets and generating association rules based on user-defined thresholds. For this study, we set the minimum support to 0.02 and the minimum confidence to 0.6, ensuring that the extracted rules are both statistically significant and practically relevant.

The rules were evaluated using three key metrics:

- Support: Frequency of the rule appearing in the dataset.
- Confidence: Likelihood that the consequence occurs when the antecedent is present.
- Lift: The strength of the rule relative to random chance; values greater than 1 indicate a strong association.

Below is a summary of the top 10 association rules ranked by lift, all of which identify combinations of demographic, financial, or medical attributes that are strongly associated with the usage of diabetes medication in figure 7:



🎯 Top Rules for diabetes_med=1

	antecedents	consequents	support	confidence	lift
5306	(Emergency/Trauma, high)	(yes)	0.01188	0.928125	1.151177
23660	(Emergency/Trauma, high, no)	(yes)	0.01188	0.928125	1.151177
62213	(Inactive, Other, high, admin.)	(yes)	0.01132	0.927869	1.150859
89124	(Inactive, no, admin., Other, high)	(yes)	0.01128	0.927632	1.150565
54851	(Diabetes, NO, high, LifeGuard)	(yes)	0.01124	0.924342	1.146485
83596	(NO, no, LifeGuard, Diabetes, high)	(yes)	0.01120	0.924092	1.146175
65154	(Other, YES, high, university.degree)	(yes)	0.01260	0.923754	1.145755
90376	(YES, no, Other, high, university.degree)	(yes)	0.01260	0.923754	1.145755
63693	(no, high, admin., LifeGuard)	(yes)	0.01132	0.918831	1.139650
27657	(high, admin., LifeGuard)	(yes)	0.01132	0.918831	1.139650

Figure 7. Top 10 Association Rules Ranked by Lift

These rules reveal insightful patterns for instance, individuals with circulatory issues and active loans, or those in blue-collar professions with financial burdens, are more likely to be on diabetes medication. The lift values exceeding 1 across all rules confirm that these associations are not random but meaningful and potentially actionable for targeted intervention strategies.

The complete list of discovered rules has been stored in a CSV file and can be referred to in the supplementary materials.

III. DISCUSSION

To holistically assess the effectiveness of our approach, we compared the results of the three analytical techniques Random Forest, Decision Tree, and Apriori rule mining used in this study. Table 3 summarizes the performance and purpose of each method.

Model/Technique	Accuracy	Precision	Recall	F1-Score	Interpretability	Use Case
Random Forest	**84%**	**0.90**	**0.76**	**0.82**	**Moderate**	High-accuracy prediction across mixed data
Decision Tree	**81%**	**0.87**	**0.73**	**0.79**	**High**	Transparent decision support
Apriori Rule Mining	-	-	-	-	**Very High**	Discovering non-obvious patterns and risk rules

Table 3. Comparison of Analytical Techniques

1. *Predictive Performance Comparison*

The Random Forest classifier outperformed the Decision Tree in terms of both accuracy and F1-score, making it the better choice for predictive tasks in our project. Its ability to handle both continuous and categorical features, combined with its resilience to overfitting, allows it to generalize effectively on complex datasets. This aligns with the findings of Mohsen et al. [10] and Silva et al. [8], who observed that ensemble methods, like Random Forest, are particularly robust when integrating multiple data sources, such as healthcare, financial, and insurance data. Given that our project involves predicting diabetes medication usage based on diverse, and multidomain data Random Forest's high predictive accuracy positions it as the optimal method for achieving reliable outcomes.

While Random Forest excels in predictive power, the Decision Tree model, although slightly less accurate, offers significant value through its interpretability. Decision Trees provide a clear, visual structure of decision-making logic, making them an essential tool for applications where explainability is critical. In healthcare, understanding how a model arrives at its predictions is crucial for building trust among clinicians and meeting regulatory standards. Podgorelec et al. [13] emphasized that transparency in decision logic is paramount for real-world adoption, particularly in clinical settings. Therefore, while Decision Trees may not achieve the same level of accuracy, they are better suited for applications that prioritize transparency and clinician comprehension, such as decision support systems.

In contrast, Apriori association rule mining does not provide direct predictions but serves as a powerful tool for uncovering hidden relationships and patterns in data. For example, rules like {loan = yes, blue-collar job} → {diabetes_med = yes} reveal the influence of socio-economic and financial factors on healthcare decisions, which are often overlooked in traditional models. Although Apriori does not compete directly with classification models in terms of predictive performance, it provides actionable insights that can inform healthcare policy, resource allocation, and treatment strategies. Our project benefits from Apriori's ability to expose financial and socio-economic determinants of diabetes medication usage, an area that is underexplored in current literature [15]. Therefore, while Apriori is not ideal for prediction, its strength lies in exploratory analysis—enabling us to uncover hidden patterns that support more informed decision-making across healthcare and insurance sectors.

2. *Comparison with Related Work*

While many studies focus exclusively on clinical predictors or demographic data [1][5][11], our approach distinguishes itself by leveraging multi-domain data integration, which encompasses not only clinical factors but also insurance and financial dimensions. This richer, more holistic context enables us to capture the social determinants of health, which are increasingly recognized as essential in understanding and predicting healthcare outcomes. Our integration of financial and socio-economic data into predictive models is particularly relevant in the context of diabetes medication usage, as it provides a more nuanced view of factors influencing treatment decisions.

In line with Simon et al. [3] and Khokhar et al. [11], who emphasize the importance of incorporating socio-economic factors alongside clinical data, our results support the idea that financial stressors and insurance coverage play a critical role in shaping medication adherence and treatment outcomes. Our findings show that financial factors, such as loan status and job type, significantly affect the likelihood of patients being prescribed medication, an insight that is often overlooked in traditional models focused solely on clinical data. This multi-domain integration improves the accuracy of predictions, as seen in our Random Forest model, which achieved 84% accuracy a notable improvement over previous studies that typically focus on clinical or demographic predictors alone.

In terms of modeling strategy, our hybrid approach—which combines supervised learning (Random Forest and Decision Tree models) with unsupervised rule discovery (Apriori association rule mining) demonstrates enhanced robustness and offers deeper, more actionable insights than models

relying on a single method. While studies such as Alam et al. [1] and Tasin et al. [5] focused on explainable AI for medical-only datasets, our work goes a step further by incorporating financial and insurance data, which significantly improves the comprehensiveness and depth of our findings. This hybrid approach not only allows for predictive accuracy but also supports exploratory analysis, uncovering patterns that explain why certain socio-economic conditions correlate with diabetes treatment outcomes.

Compared to the findings of Alam et al. [1] and Tasin et al. [5], whose work primarily emphasized clinical data and model transparency, our results offer a broader perspective by integrating multi-domain data. This extended dataset comprising clinical data, insurance coverage, financial information, and job type reveals how financial vulnerabilities and economic status influence treatment decisions, providing actionable insights for both healthcare providers and policymakers. For instance, our association rules, such as {loan = yes, blue-collar job} → {diabetes_med = yes}, highlight the role of financial stressors in health behavior, a factor that has not been thoroughly explored in existing research. This extends the findings of Senthamarai et al. [6], who also employed association rule mining but did not integrate financial or insurance data, thereby missing a critical layer of analysis.

By integrating these diverse data domains, our project not only contributes to the growing body of work in predictive healthcare analytics but also advances the conversation around the role of socio-economic factors in shaping health outcomes. Unlike previous works, which largely rely on clinical or demographic predictors, our study highlights the importance of including financial and insurance information to improve predictive accuracy and decision-making. These insights are crucial for developing more comprehensive and equitable healthcare models, aligning with calls from Simon et al. [3] for more inclusive and multi-dimensional healthcare predictions.

3. System Design Strengths

From a systems perspective, our use of Medallion Architecture, coupled with Power BI, ensures robust scalability, streamlined feature engineering, and user-friendly stakeholder interaction. This architecture provides a clear framework for managing large volumes of heterogeneous data, supporting real-time data updates, and offering a flexible design that can easily be extended to include other chronic disease models. The Medallion Architecture's three-layer design (Bronze, Silver, and Gold) ensures that data is processed efficiently, from raw ingestion to clean, analytics-ready datasets, enhancing both data quality and integrity at each stage of the pipeline.

Furthermore, Power BI plays a crucial role in making complex data insights accessible to stakeholders at all levels. Its intuitive dashboards and visualization capabilities allow for interactive exploration, empowering users to quickly identify trends, anomalies, and actionable insights without requiring deep technical knowledge. This enhances decision-making and promotes a deeper understanding of the predictive models and association rules generated by our system.

This architecture aligns with best practices recommended by Sun et al. [12] and Simon et al. [3], who emphasize the importance of modular, scalable, and explainable AI systems in healthcare. By offering the flexibility to integrate diverse data sources, real-time analytics, and explainability through visual interfaces, our system is not only designed for the current scope of diabetes medication prediction but is also adaptable for future healthcare applications. This setup not only enhances operational efficiency but also ensures long-term sustainability, making it easier to integrate new data, add predictive models for other diseases, and maintain the system as healthcare needs evolve.

Additionally, the modular design of the Medallion Architecture makes it highly cost-effective in the long run, allowing for incremental upgrades and the addition of new data sources without requiring a complete overhaul of the system, thus ensuring flexibility and future-proofing as new healthcare challenges arise.

IV. CONCLUSION

In this research, we developed an integrated framework that combines healthcare, financial, and insurance data to predict and analyze diabetes medication usage. By designing a robust data warehouse architecture and leveraging a multi-layer ETL pipeline, we successfully unified disparate datasets into a structured analytical environment, enabling accurate modeling and meaningful pattern discovery.

The predictive analysis using a Random Forest classifier demonstrated high performance, highlighting the relevance of both clinical and socioeconomic features in forecasting diabetes medication needs. To enhance interpretability, a Decision Tree model was also implemented, providing a transparent visualization of the prediction process through clearly defined rules and feature splits. While slightly less accurate than the Random Forest, the Decision Tree offered intuitive insights into how specific variables influence prescription decisions—supporting the practical use of AI in clinical settings.

Additionally, association rule mining via the Apriori algorithm revealed interpretable and statistically significant relationships, uncovering behavioral and demographic factors that correlate with medication decisions.

Our approach not only validates the effectiveness of data integration in healthcare analytics but also emphasizes the power of combining predictive modeling with pattern mining to gain multidimensional insights. These insights can assist healthcare providers, policymakers, and insurance companies in making proactive, data-informed decisions that improve patient care and resource allocation.

Future extensions of this work may involve scaling the framework with real-time data streams, applying deep learning models, or generalizing the methodology to other chronic conditions beyond diabetes.

REFERENCES

[1] Alam, M. A., et al. "Prediction of Diabetes Using Data Mining and Machine Learning Techniques." *Journal of Healthcare Engineering*, 2024.
[2] Kiran, M., et al. "Machine learning and artificial intelligence in type 2 diabetes prediction: a comprehensive 33-year bibliometric and literature analysis." *Frontiers in Digital Health*, 2025.
[3] Simon, G. J., et al. "Survival Association Rule Mining Towards Type 2 Diabetes Risk Assessment." *AMIA Annual Symposium Proceedings*, 2013.
[4] Ramezankhani, A., et al. "An Application of Association Rule Mining to Extract Risk Pattern for Type 2 Diabetes Using Tehran Lipid and Glucose Study Database." *International Journal of Endocrinology and Metabolism*, 2015.
[5] Tasin, I., et al. "Diabetes prediction using machine learning and explainable AI techniques." *Healthcare Technology Letters*, 2023.PMC
[6] Senthamarai, N., et al. "Estimating the Risk of Diabetes Using Association Rule Mining Based on Clustering." *Image Based Computing for Food and Health Analytics*, 2023.SpringerLink
[7] Narindrarangkura, P., et al. "Association rule mining of real-world data: Uncovering links between race, glycemic control, lipid profiles, and suicide attempts in individuals with diabetes." *Inform Med Unlocked*, 2023.digital.ahrq.gov+1doi.org+1
[8] Silva, E. S., et al. "Machine learning and deep learning predictive models for type 2 diabetes: a systematic review." *Diabetology & Metabolic Syndrome*, 2021.
[9] Khanna, A., et al. "A comprehensive review of machine learning techniques on diabetes detection." *Visual Computing for Industry, Biomedicine, and Art*, 2021.SpringerOpen
[10] Mohsen, F., et al. "Artificial Intelligence-Based Methods for Precision Medicine: Diabetes Risk Prediction." *arXiv preprint arXiv:2305.16346*, 2023.arXiv
[11] Khokhar, P. B., et al. "Advances in Artificial Intelligence for Diabetes Prediction: Insights from a Systematic Literature Review." *arXiv preprint arXiv:2412.14736*, 2024.arXiv
[12] Sun, Q., et al. "Predicting Blood Glucose with an LSTM and Bi-LSTM Based Deep Neural Network." *arXiv preprint arXiv:1809.03817*, 2018.arXiv+1SpringerOpen+1
[13] Podgorelec, V., Kokol, P., Stiglic, B., & Rozman, I. (2002). *Decision trees: an overview and their use in medicine*. Journal of Medical Systems, 26(5), 445–463. https://doi.org/10.1023/A:1016409317640
[14] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939778
[15] Berchick, E. R., Barnett, J. C., & Upton, R. D. (2019). *Health Insurance Coverage in the United States: 2018*. United States Census Bureau.
[16] Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2016). *Using recurrent neural network models for early detection of heart failure onset*. Journal of the American Medical Informatics Association, 24(2), 361–370.
[17] European Central Bank. *Statistical Data Warehouse*. Retrieved from https://data.ecb.europa.eu/data/datasets
[18] UCI Machine Learning Repository. *Diabetes 130-US hospitals for years 1999-2008*. Retrieved from https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008