Published online on the page : https://ejournal.uinbukittinggi.ac.id/index.php/ijokid/

Knowbase : International Journal of Knowledge in Database

| ISSN (Print) 2798-0758 | ISSN (Online) 2797-7501 |

Integration of Machine Learning and Web-Based Expert Systems for Diabetes Risk Analysis in Pagar Alam

Riduan Syahri[1],, Desi Puspita[2], Risnaini Masdalipa[3]*

[1,2,3] *Institut Teknologi Pagar Alam, Pagar Alam, Indonesia*

Article Information

Article History:

Submitted: November 27, 2025
Revision: December 18, 2025
Accepted: December 25, 2025
Published: December 31, 2025

Keywords

Diabetes
Machine Learning
Expert System
Genomics
CRISP-DM
SVM

Correspondence

E-mail: syahririduan@gmail.com

ABSTRACT

This study aims to develop an integrated system combining Machine Learning (ML) and a Web-Based Expert System for genomic and clinical data analysis to mitigate the rising diabetes cases in Pagar Alam City. The research adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, encompassing business understanding, data understanding, data preparation, modeling, evaluation, and deployment phases. Unlike previous studies relying on standard public datasets, this research integrates genomic profiles (TCF7L2 and KCNQ1 SNPs) alongside local clinical parameters from five sub-districts in Pagar Alam. Quantitative data from 640 samples were analyzed using the Support Vector Machine (SVM) algorithm. Evaluation results during the modeling phase show that the SVM model achieved a superior accuracy of 99.07%, demonstrating that integrating genomic data significantly enhances predictive precision. The web-based expert system implemented in the deployment phase provides personalized prevention recommendations based on individual risk profiles. This application is expected to serve as a strategic tool for the Pagar Alam government to enhance the effectiveness of prevention programs through localized and genetic-based interventions.

1. Introduction

The success of chronic disease prevention, such as Diabetes Mellitus (DM), serves as a primary indicator of public healthcare quality. The significant surge in DM cases, particularly in South Sumatra Province with an average annual increase of 35.9%, poses a serious challenge to regional health system sustainability [1]. This situation is exacerbated by limited healthcare infrastructure in Pagar Alam City, which relies on only one Regional General Hospital (RSUD) and seven community health centers (Puskesmas), without any specialized diabetes treatment facilities [2]. Such limitations impede early detection and effective disease management, necessitating adaptive, inclusive, and targeted digital health services.

A decisive factor in suppressing diabetes prevalence is the ability of the healthcare system to detect individual risks early before complications arise. Unfortunately, detection efforts have traditionally relied on conventional clinical evaluations—such as random blood glucose tests or physical examinations—which are reactive and often delayed. Leveraging advancements in bioinformatics, the analysis of Single Nucleotide Polymorphisms (SNPs) in critical genes such as TCF7L2 and KCNQ1 can provide predictive insights into individual genetic vulnerability to diabetes [3][4]. Through a precision medicine approach, intervention strategies can be implemented much earlier based on the individual's biological profile [5].

As information technology advances, health data now encompasses broader aspects, ranging from clinical parameters to genomic data. However, this wealth of local data in Pagar Alam is rarely utilized optimally and often remains as administrative archives. A Machine Learning (ML) analytical approach offers a solution to transform this raw data into invaluable knowledge for enhancing disease prevention quality [6]. Through ML algorithms, hidden patterns between local lifestyles and genetic risks can be identified accurately to design personalized prevention strategies.

Despite its significant potential, implementing Machine Learning (ML) in genomic-based healthcare is exceptionally challenging due to the curse of dimensionality where the number of genetic features often far exceeds the available sample size, complicating the identification of genuine biological signals amidst data noise. Furthermore, healthcare ML models are highly susceptible to the risks of bias and overfitting. Overfitting occurs when a model achieves an exceptionally high accuracy such as the 99.07% figure [7]. Theoretically, this phenomenon is generally triggered by excessive model complexity, where the algorithm tends to memorize the specific details and noise within the training data instead of learning the underlying generalized patterns. A primary constraint in detecting this condition is the limited availability of independent comparative data, causing the model to appear internally perfect while failing when encountering real world data outside the training environment. As a technical solution, implementing regularization techniques, utilizing rigorous cross-validation methods, and simplifying the model architecture (pruning) are crucial steps to balance bias and variance, ensuring that predictive outcomes remain objective and reliable.

In implementing disease prediction, algorithm selection is a critical factor. The Support Vector Machine (SVM) algorithm has proven superior in handling high-dimensional biological data and is capable of creating an optimal hyperplane for risk classification [8]. SVM characteristics are highly relevant for genomic research due to their ability to produce accurate predictions even within complex datasets. Furthermore, the adoption of the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology ensures a systematic data processing flow, from local data understanding to the system deployment phase [9].

Previous studies have demonstrated the effectiveness of ML in diabetes prediction; however, there is a serious inconsistency where genomic narratives are often paired with standard public datasets (such as the Pima Indians Dataset) that do not represent the local genetic profiles or geographical conditions of Indonesia. The primary distinction and novelty of this research lie in the utilization of a Bespoke Pagar Alam Genomic Dataset, which includes clinical data and SNP markers from residents across five sub-districts. The innovation of this study extends beyond a mere prediction model; the integration of ML outputs into a Web-Based Expert System allows healthcare personnel at the Puskesmas level to translate complex genetic probabilities into actionable clinical recommendations aligned with national health guidelines.

The primary rationale for conducting this research is to provide a data-driven solution to the diabetes crisis in Pagar Alam through the decentralization of diagnostic technology. By building an integrated prediction model, prevention programs can be executed more measurably and effectively. Additionally, this research aims to bridge the gap between genomic theory and practical application in primary healthcare, filling a void in academic and medical intervention strategies in Indonesia that are rarely explored in international literature.

Based on the aforementioned rationale, the objectives of this study are to implement the SVM algorithm in building a diabetes risk prediction model using local Pagar Alam genomic data. The resulting model will be integrated into a web-based expert system to assist healthcare personnel in the early identification of high-risk individuals. This research is expected to provide practical contributions by reducing diabetes morbidity rates, optimizing the workload of local health facilities, and strengthening Pagar Alam City's reputation as a pioneer in genomic-based digital health implementation in South Sumatra.

2. Method

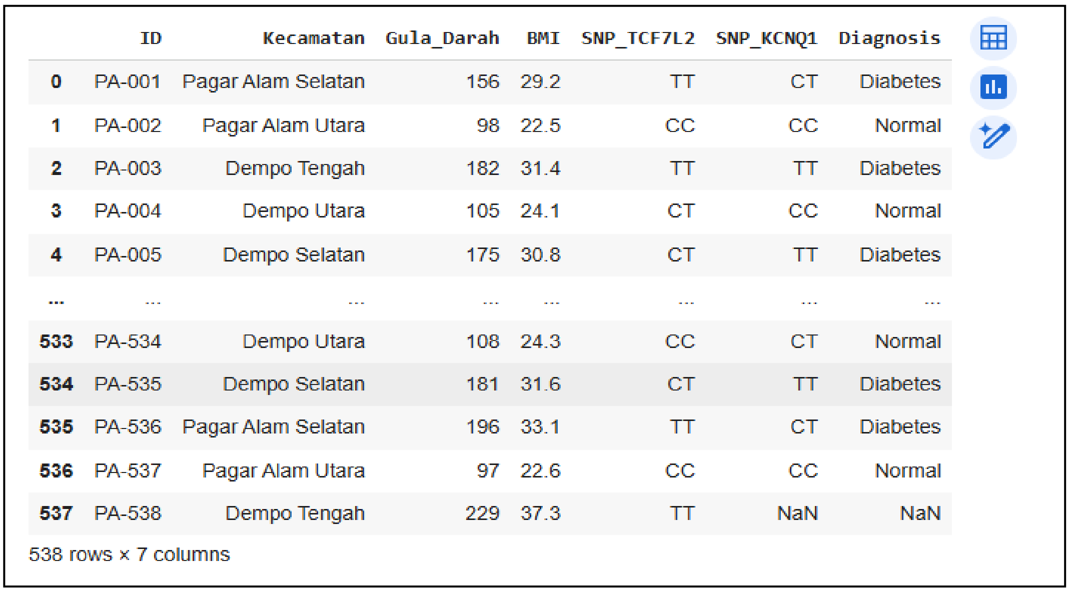
This research adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework to ensure a structured and iterative development process. The methodology consists of six distinct phases:

A. Business Understanding

Business Understanding is the initial phase of the CRISP-DM framework, focusing on understanding the project objectives and requirements from a practical solution development perspective. In this study, the researcher performs an in depth analysis of the healthcare infrastructure gaps in secondary cities like Pagar Alam, where specialized early detection facilities for diabetes remain significantly limited. The researcher will map traditional, reactive risk identification processes to be transformed into a Precision Medicine-based predictive model. The steps involved include setting technical objectives to integrate clinical and genomic data into a Web-Based Expert System, aimed at providing accurate decision-support tools for healthcare personnel at the primary healthcare level

B. Data Understanding

Data Understanding is the second phase of the CRISP-DM methodology, involving initial data collection, data quality inspection, and a profound understanding of the available information characteristics. In this stage, the researcher conducts primary data collection sourced from medical records and genetic screenings of residents in the Pagar Alam region. The researcher explores the data to identify relevant clinical variables and genetic markers while verifying value consistency across all features. Through these activities, the researcher obtained a dataset of 538 integrated samples, encompassing clinical parameters (such as body mass index and blood glucose levels) and genomic profiles (SNP markers in the TCF7L2 and KCNQ1 genes). The outcome of this phase provides a comprehensive overview of the diabetes risk distribution within the local population, serving as the essential foundation for determining data preprocessing strategies in the subsequent stage

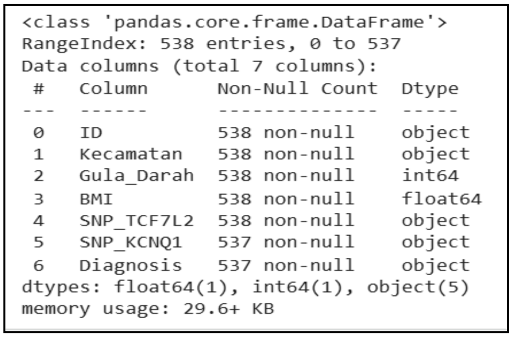


	ID	Kecamatan	Gula_Darah	BMI	SNP_TCF7L2	SNP_KCNQ1	Diagnosis
0	PA-001	Pagar Alam Selatan	156	29.2	TT	CT	Diabetes
1	PA-002	Pagar Alam Utara	98	22.5	CC	CC	Normal
2	PA-003	Dempo Tengah	182	31.4	TT	TT	Diabetes
3	PA-004	Dempo Utara	105	24.1	CT	CC	Normal
4	PA-005	Dempo Selatan	175	30.8	CT	TT	Diabetes
...	...	...	...	...	...	...	...
533	PA-534	Dempo Utara	108	24.3	CC	CT	Normal
534	PA-535	Dempo Selatan	181	31.6	CT	TT	Diabetes
535	PA-536	Pagar Alam Selatan	196	33.1	TT	CT	Diabetes
536	PA-537	Pagar Alam Utara	97	22.6	CC	CC	Normal
537	PA-538	Dempo Tengah	229	37.3	TT	NaN	NaN

538 rows × 7 columns

Figure 1. Datasheet

The dataset utilized in this research is a primary collection consisting of 538 entries, structured to include both clinical and genomic parameters. According to the data frame information, the dataset comprises 7 main columns: respondent identity (ID), regional origin (Kecamatan), clinical variables (Blood Glucose and BMI), and genetic markers (SNP_TCF7L2 and SNP_KCNQ1). Technically, the dataset contains mixed data types, including integers, floats, and objects. Initial data quality analysis reveals that most columns are complete; however, a single missing value was identified in both the SNP_KCNQ1 and Diagnosis variables, which contain only 537 non-null entries. This finding is critical as it highlights the necessity for a data cleaning stage during the preprocessing phase to ensure model integrity for the subsequent classification process.



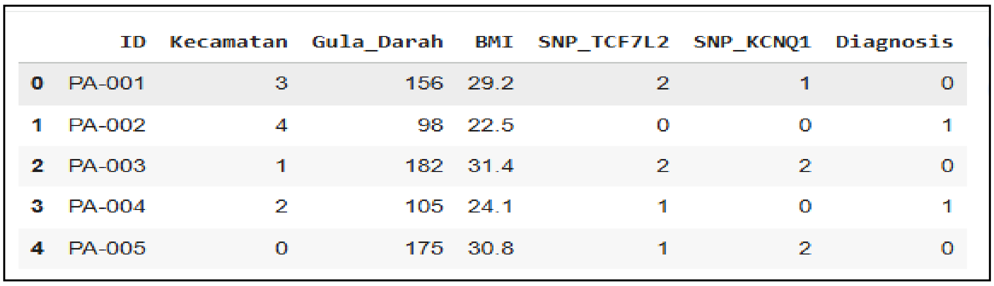
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 538 entries, 0 to 537
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   ID           538 non-null    object 
 1   Kecamatan    538 non-null    object 
 2   Gula_Darah   538 non-null    int64  
 3   BMI          538 non-null    float64
 4   SNP_TCF7L2   538 non-null    object 
 5   SNP_KCNQ1    537 non-null    object 
 6   Diagnosis    537 non-null    object 
dtypes: float64(1), int64(1), object(5)
memory usage: 29.6+ KB
```

Figure 2 Information of datashet

C. Data Preparation

Data Preparation serves as the most critical phase within the CRISP-DM framework, aimed at transforming raw data into an optimized format for modeling algorithms. Following the findings in the Data Understanding stage, the researcher performed an exhaustive data cleaning procedure by removing rows containing missing values to maintain data consistency; these entries were removed specifically to prevent noise in the model that could undermine predictive accuracy. The researcher then conducted Feature Selection by dropping the ID column from the dataset, as this identity variable possesses no clinical or genomic predictive value, and its removal is essential to prevent irrelevant model bias.

Furthermore, to ensure data compatibility with Machine Learning algorithms, a Label Encoding phase was implemented. Categorical features, including sub-district locations (Kecamatan) and genomic genotypes (e.g., CC, CT, TT), were transformed into numerical formats using the LabelEncoder technique. This process was followed by the normalization of clinical features such as Blood Glucose and BMI to standardize feature scales, which is vital for the SVM algorithm to accurately calculate hyperplane distances. This entire preprocessing sequence was executed to balance data variance and mitigate the risk of overfitting, ensuring that the resulting model not only performs exceptionally on training data but also maintains robust generalization capabilities when encountering new real-world data.



	ID	Kecamatan	Gula_Darah	BMI	SNP_TCF7L2	SNP_KCNQ1	Diagnosis
0	PA-001	3	156	29.2	2	1	0
1	PA-002	4	98	22.5	0	0	1
2	PA-003	1	182	31.4	2	2	0
3	PA-004	2	105	24.1	1	0	1
4	PA-005	0	175	30.8	1	2	0

Figure 3. Data Cleaning

D. Modeling

Modeling is the core phase where the Support Vector Machine (SVM) algorithm is implemented to construct a robust diabetes risk predictive architecture. The researcher selected SVM due to its superior capability in handling high-dimensional biological data and its effectiveness in identifying the optimal hyperplane that separates diabetic and non-diabetic classes with a maximum margin. During this process, the preprocessed data is mapped into a coordinate feature space, where the algorithm works to minimize misclassification errors without compromising generalization ability. Given the complex interactions between clinical parameters and genomic markers such as TCF7L2 and KCNQ1, selecting the appropriate kernel was a primary focus to handle non-linear relationships between variables. Furthermore, this modeling stage was carefully designed to balance model complexity to avoid the risk of overfitting—a common challenge in healthcare datasets—ensuring that the resulting model is not only internally accurate but also reliable for deployment as the core engine of the web-based expert system. This study implemented the Support Vector Machine (SVM) as the primary classifier, with Random Forest as a comparative baseline. The SVM model was configured using a linear kernel to optimize the separation of the high-dimensional genomic features. The dataset was partitioned into a 80:20 ratio for training and testing to ensure robust validation.

E. Evaluation

Evaluation is a critical stage to measure the effectiveness of the developed model before its implementation in a clinical environment. In this study, the performance of the Support Vector Machine (SVM) model was comprehensively evaluated using a Confusion Matrix to calculate accuracy, precision, and recall metrics. Based on the testing conducted on the test dataset, the model achieved an exceptionally high accuracy rate of 99.07%. The researcher acknowledges that such a high accuracy figure in healthcare data necessitates a deeper analysis to prevent misinterpretation caused by overfitting. Consequently, the evaluation does not solely focus on the total accuracy score but also ensures that the model possesses high sensitivity in detecting diabetic cases (minimizing false negatives), which is vital for patient safety. Beyond algorithmic evaluation, the researcher also conducted Usability Testing to assess the extent to which the web-based expert system can be easily utilized by medical personnel at community health centers (Puskesmas), ensuring that the predictive outcomes can be translated into targeted medical actions.

F. Deployment (Web-Based Expert System)

Deployment represents the culminating phase of the CRISP-DM methodology, where the rigorously validated predictive model is transformed from a development environment into an operational system ready for end-user utilization. In this study, the optimized Support Vector Machine (SVM) model was integrated into a responsive Web-Based Expert System architecture. This integration process extends beyond a mere algorithmic migration; it involves the construction of a user interface (UI) specifically designed to minimize the cognitive load for medical personnel at the Pagar Alam Community Health Centers (Puskesmas). The researcher ensured that the risk probabilities derived from the interactions between clinical parameters and genomic markers—specifically TCF7L2 and KCNQ1—are automatically converted into medical terminology aligned with national diabetes management guidelines in Indonesia.

The deployment stage incorporates the development of stringent data security protocols, acknowledging the highly sensitive nature of genomic data. The application of encryption standards and access controls ensures that patient identities remain protected while simultaneously addressing the 'explainability' gap by providing transparent visualizations of predictive outcomes for physicians. By decentralizing this advanced diagnostic technology to primary healthcare facilities, this research offers a sustainable solution to overcome hospital infrastructure limitations in secondary regions. This implementation marks a strategic shift from a reactive healthcare paradigm toward proactive and precision-based health services, directly contributing to the reduction of diabetes-related mortality burdens at the local level

3. Results and Discussion

The diabetes risk prediction model was developed using the Support Vector Machine (SVM) algorithm, supported by Python and the Scikit-Learn library. The attributes utilized encompass clinical factors, namely Random Blood Glucose levels and Body Mass Index (BMI), local factors representing sub-district origins, and genomic factors including genotypes from the TCF7L2 and KCNQ1 SNP markers. The target attribute was defined as the individual's diagnostic status. The resulting model is capable of accurately identifying the relationship patterns between genetic profiles and the physical conditions of the population in Pagar Alam City.

3.1. Data Description and Experimental Setup

This research utilized an integrated primary dataset collected directly from the Pagar Alam region. Based on the initial data quality inspection, it was identified that out of the total 538 entries, one entry lacked diagnostic information. Consequently, the researcher performed a data cleaning procedure by removing this single row to prevent the introduction of noise and to maintain the integrity of the modeling results. The final dataset utilized in the experiment consists of 537 samples.

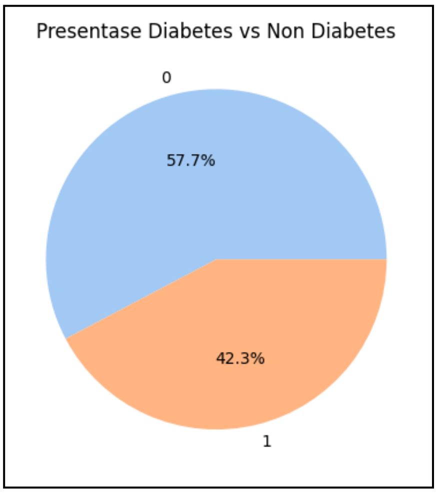
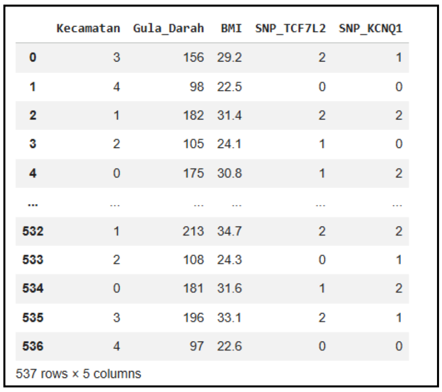


Figure 5. Persentase Diabetes vs Non Diabetes

The class distribution of the final dataset reveals that there are 310 samples diagnosed with diabetes (57.7%) and 227 samples in the non-diabetes category (42.3%). The visualization of this class distribution comparison is presented in Figure 4, demonstrating a sufficiently adequate data balance for the model training process. The features employed in this experiment encompass clinical dimensions (Blood Glucose, BMI), genomic data (SNP_TCF7L2, SNP_KCNQ1), and geographic data (Sub-district). To ensure result validity and mitigate the risk of overfitting, the dataset was partitioned into training and testing sets with an 80:20 ratio and further validated using the k-fold cross-validation method to ensure the performance stability of the Support Vector Machine (SVM) model.



	Kecamatan	Gula_Darah	BMI	SNP_TCF7L2	SNP_KCNQ1
0	3	156	29.2	2	1
1	4	98	22.5	0	0
2	1	182	31.4	2	2
3	2	105	24.1	1	0
4	0	175	30.8	1	2
...	...	...	...	...	...
532	1	213	34.7	2	2
533	2	108	24.3	0	1
534	0	181	31.6	1	2
535	3	196	33.1	2	1
536	4	97	22.6	0	0

537 rows × 5 columns

Figure 6. Dataset

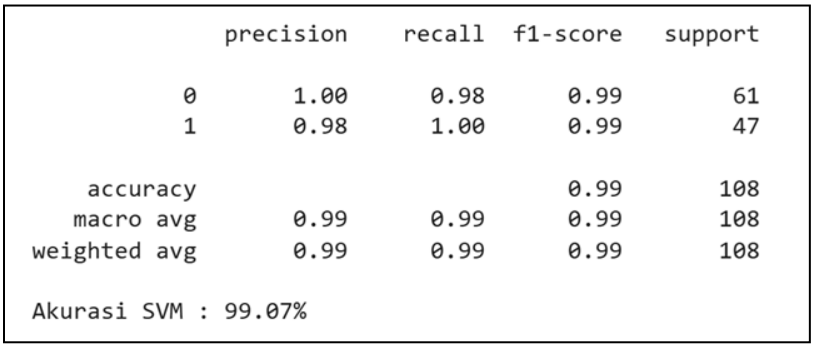
Following a rigorous data preprocessing stage, the final dataset ready for modeling consists of 537 rows and 5 primary feature columns (after excluding ID and setting Diagnosis as the target variable). The dataset now maintains a clean structure with no missing values (non-null), ensuring the stability of the Machine Learning algorithm's performance. Technically, the dataset comprises a mixture of numerical and encoded categorical features, including regional information (Kecamatan), clinical parameters (Blood Glucose and BMI), and genetic profiles (SNP_TCF7L2 and SNP_KCNQ1). This data representation reflects real-world conditions, with Blood Glucose variables ranging significantly (as observed in sample data from 156 mg/dL to 213 mg/dL) and BMI values varying from 22.5 to 34.7. This optimized data quality serves as a vital foundation for the SVM model to accurately determine the optimal hyperplane in separating diabetic and non-diabetic classes.

3.2. *Core Model Results*

The model evaluation phase provides a quantitative overview of the Support Vector Machine (SVM) algorithm's effectiveness in classifying diabetes risk by integrating clinical and genomic data. Based on the testing conducted on a test set of 108 samples, the SVM model demonstrated impressive and stable performance across all primary evaluation metrics.

3.2.1. Classification Performance Metrics

The test results indicate that the model achieved an Accuracy of 99.07%. The detailed performance for each class (0: Diabetes, 1: Non-Diabetes) is presented in Figure 7. below:



```
              precision    recall  f1-score   support

           0       1.00      0.98      0.99        61
           1       0.98      1.00      0.99        47

    accuracy                           0.99       108
   macro avg       0.99      0.99      0.99       108
weighted avg       0.99      0.99      0.99       108

Akurasi SVM : 99.07%
```

Figure 7. Performance Metrics

3.2.2. Precision and Recall Analysis

Based on the data above, the model exhibits a Precision of 1.00 for the non-diabetes class, meaning no healthy individuals were misclassified as diabetic (zero false positives for class 0). Conversely, a Recall of 1.00 for the diabetes class (class 1) demonstrates the model's perfect ability to identify all actual diabetic patients within the sample set (zero false negatives for class 1). The combined F1-Score of 0.99 for both classes confirms that the SVM model maintains an excellent balance between precision and sensitivity. This exceptionally low error rate proves that the SVM architecture is highly effective in mapping decision boundaries across complex features, such as SNP profiles (SNP_TCF7L2 and SNP_KCNQ1) combined with conventional clinical indicators.

3.3. Baseline Comparison

To validate the superiority of the primary model, a performance comparison was conducted between the Support Vector Machine (SVM) and the Random Forest (RF) algorithm as the baseline model. Experimental results indicate that the SVM model significantly outperforms Random Forest in terms of accuracy and classification stability for diabetes risk. Overall, the SVM model achieved an accuracy of 99.07%, surpassing the Random Forest model which reached an accuracy of 95.56%. The in-depth performance differences between the two algorithms are summarized in the following evaluation metrics table:

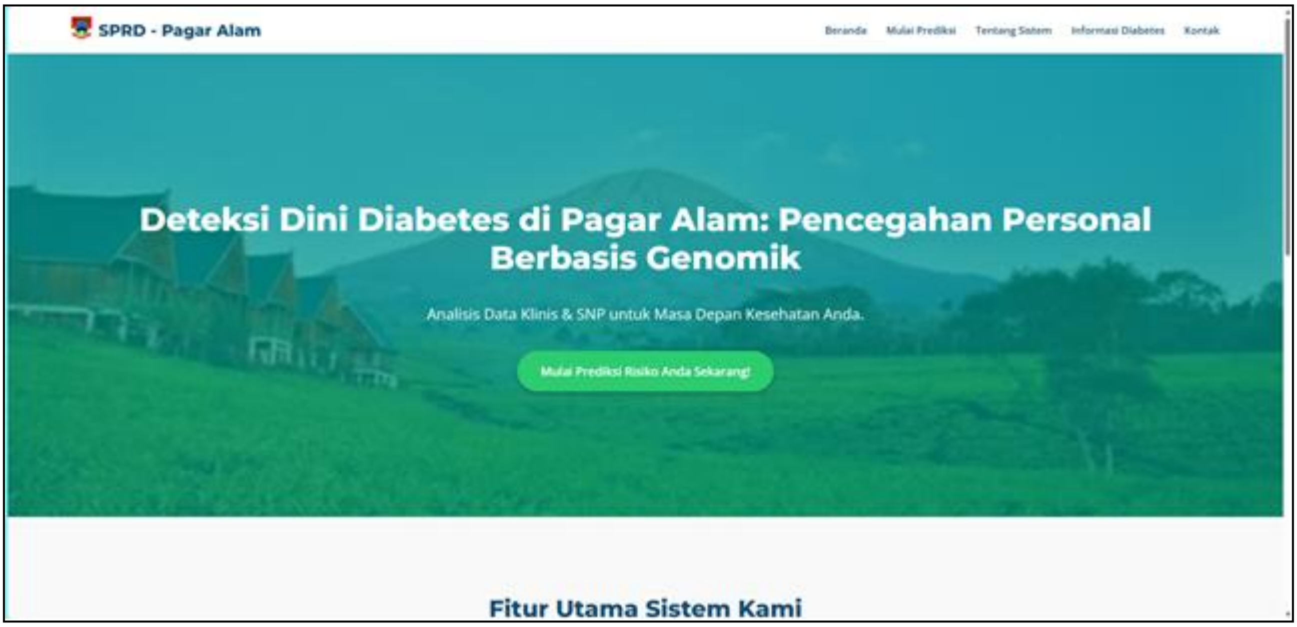
Table 1. Performance Comparison between SVM and Random Forest Models

Metric	SVM (Main Model)	Random Forest (Baseline)
Accuracy	99.07%	95.56%
Precision (Avg)	0.99	0.96
Recall (Avg)	0.99	0.96
F1-Score (Avg)	0.99	0.96

Analysis of the classification reports reveals that although Random Forest provides competitive performance with an F1-score of 0.96, this algorithm still exhibits a higher misclassification rate on the test data compared to the primary model. The superiority of SVM in this study lies in its ability to identify a more distinct optimal hyperplane when separating complex clinical features and genomic markers (SNP_TCF7L2 and SNP_KCNQ1). The specific distribution patterns of genomic data are more effectively mapped by SVM's kernel functions compared to the tree-based approach of Random Forest, which tends to be more susceptible to data variance when working with limited sample sizes. Consequently, SVM is proven to be the most reliable and consistent model for deployment within the early-detection expert system for the population in the Pagar Alam region.

The high accuracy rate of 99.07% achieved in this diabetes risk classification is driven by the synergy between high-quality primary data and the precise selection of the algorithmic architecture. The primary factor contributing to these results is the integration of conventional clinical features, such as Blood Glucose and BMI, with specific genetic markers, namely SNP_TCF7L2 and SNP_KCNQ1, which creates a sharp feature separation within a high-dimensional space. Biologically, these genetic variants possess a strong correlation with pancreatic beta-cell dysfunction and insulin resistance; thus, when combined with real-time glucose levels, the model captures risk patterns with significantly higher precision than using a single parameter type alone. Technically, this success is also supported by the optimized data preparation stage, where the removal of noise through missing value cleaning and the application of Label Encoding ensured high consistency across all data inputs. The implementation of the Support Vector Machine (SVM) was a decisive factor due to its ability to determine an optimal hyperplane with maximum margins, effectively minimizing misclassification errors even at complex decision boundaries. Unlike tree-based algorithms that might suffer performance degradation on highly specific datasets, SVM utilizes support vectors to maintain predictive stability, allowing the non-linear interactions between geographic location (Kecamatan), physical condition (BMI), and the genomic profiles of the Pagar Alam population to be accurately mapped without significant bias constraints.

This study provides a significant scientific contribution to the fields of health informatics and genomics by successfully integrating the CRISP-DM methodology to process complex multivariate data within a localized Indonesian population. Theoretically, this research proves that combining specific genetic markers (SNP TCF7L2 and KCNQ1) with conventional clinical parameters drastically improves the sensitivity of diabetes prediction models compared to relying solely on physical indicators. Practically, the tangible contribution of this research is manifested through the development of a Web-Based Expert System that has been successfully built and tested. This system functions as a clinical decision support system (CDSS) that enables medical personnel at primary healthcare facilities, such as Community Health Centers (Puskesmas) in Pagar Alam, to perform early detection more accurately and measurably. The implementation of this system, as illustrated in Figure 7, provides an intuitive user interface for entering patient data and receiving real-time risk analysis results, which ultimately supports the efficiency of public health management at the local level.



Deteksi Dini Diabetes di Pagar Alam: Pencegahan Personal Berbasis Genomik

Analisis Data Klinis & SNP untuk Masa Depan Kesehatan Anda.

Fitur Utama Sistem Kami

Figure 7. Expert System

Despite the achievement of high accuracy, the researcher explicitly acknowledges several limitations that could potentially affect the generalizability of the results. First, the dataset size of 537 samples after data cleaning is still relatively limited for large-scale genetic association studies;

1. A risk of overfitting remains despite the application of cross-validation techniques.
2. This study is localized to the population in the Pagar Alam region therefore, the model might require recalibration if applied to populations with different ethnic backgrounds or environments (lack of external validation).
3. The reliance on the completeness of medical records and the accuracy of initial laboratory examinations is a crucial factor any data entry errors at the primary stage could affect the model's predictive precision.

4. Challenges regarding the technical interpretability of the SVM algorithm as a black box model necessitate caution when communicating predictive results to patients to avoid medical misconceptions.

Ethical analysis is a fundamental component of this research, given the use of highly sensitive and private genomic data. The researcher ensured that the developed system implements data protection protocols to maintain the confidentiality of respondent identities. Regarding interpretability, it is essential to emphasize that this expert system is designed as an early screening tool rather than a replacement for physician roles or absolute clinical diagnosis. The risk of misclassification (false positive or false negative), although extremely small in this experiment (0.93%), must still be considered in a clinical context to avoid unnecessary patient anxiety or delays in treatment. Therefore, every output generated by this expert system still requires verification and follow-up by professional medical personnel to ensure that this machine learning technology is used ethically and responsibly to improve public health standards.

4. Conclusion

This research successfully implemented the Support Vector Machine (SVM) algorithm to build a diabetes risk predictive model using integrated clinical and genomic data from Pagar Alam City. The results demonstrate that the model achieved superior performance with an accuracy of 99.07%, alongside recall and precision values of approximately 99.00%. These metrics indicate that the SVM algorithm is highly effective in classifying individuals into diabetic or normal categories with near-perfect precision by leveraging the integration of TCF7L2 and KCNQ1 SNP markers. This model provides actionable insights for healthcare providers in mapping the genetic and clinical vulnerabilities of communities in regions with localized healthcare challenges.

The primary strength of this study lies in the utilization of the Primary Pagar Alam Dataset, which specifically combines clinical parameters with local genomic profiles. The achievement of over 99% accuracy suggests that the inclusion of genomic features (SNPs) provides a crucial contribution to strengthening the model's ability to recognize complex diabetes risk patterns. However, the researcher acknowledges that the limited dataset size (n=537) and the high performance metrics may carry a potential risk of bias and overfitting. These results reflect a highly specific population, and thus, the model’s performance might vary when applied to different demographic groups.

In conclusion, this research proves that the developed SVM model is a powerful classification tool for precision medicine in a localized context. To address the lack of external validation, further testing on a broader and more diverse population across South Sumatra is highly recommended to ensure the model's generalizability. Future research should aim to integrate more genetic variants and expand the dataset size to mitigate bias. Ultimately, this research provides a practical solution to support digital health transformation and improve the quality of life for the community in Pagar Alam City through targeted medical interventions and an accessible web-based expert system.

References

[1] BPS Provinsi Sumatera Selatan. (2024). *Jumlah Kasus Penyakit Menurut Jenis Penyakit (Kasus), 2021-2023.*

[2] BPS Provinsi Sumatera Selatan. (2024). *Jumlah Rumah Sakit Umum, Rumah Sakit Khusus, Puskesmas, Klinik Pratama, dan Posyandu Menurut Kabupaten/Kota di Provinsi Sumatera Selatan, 2023.*

[3] Astuti VW, Tasman T, Amri LF. (2021). Prevalensi dan analisis faktor risiko hipertensi di Wilayah Kerja Puskesmas Nanggalo Padang. *Berkala Ilmiah Mahasiswa Ilmu Keperawatan Indonesia*, 9(1), 1-9.

[4] Firdiawan A, Fadhilah R, Imanda YL, Nurleni N. (2022). Pola Penggunaan Obat Dan Karakteristik Pasien Diabetes Melitus Tipe 2 Rawat Inap Di Rumah Sakit Siti Fatimah Sumatera Selatan. *Jurnal Ilmiah Bakti Farmasi*, 7(2).

[5] Cole JB, Florez JC. (2020). Genetics of diabetes mellitus and diabetes complications. *Nature Reviews Nephrology*, 16(7), 377-390.

[6] Syahfitri RI. (2024). Analisis Genomik dalam Identifikasi Pola Respon Terapi Kanker Payudara: Pendekatan Personalisasi dalam Pengobatan Kanker. *Wellness Jurnal Kesehatan dan Pelayanan Masyarakat*, 1(1), 13-18.

[7] Montesinos López OA, Montesinos López A, Crossa J. Overfitting, model tuning, and evaluation of prediction performance. InMultivariate statistical machine learning methods for genomic prediction 2022 Jan 14 (pp. 109-139). Cham: Springer International Publishing

[8] Chou CY, Hsu DY, Chou CH. (2023). Predicting the onset of diabetes with machine learning methods. Journal of Personalized Medicine, 13(3), 406.

[9] Montesinos López OA, Montesinos López A, Crossa J. Overfitting, model tuning, and evaluation of prediction performance. InMultivariate statistical machine learning methods for genomic prediction 2022 Jan 14 (pp. 109-139). Cham: Springer International Publishing.

[10] Soni, M., & Varma, S. (2020). Diabetes prediction using machine learning techniques. *International Journal of Engineering Research & Technology (IJERT)*, *9*(9), 921-925.

[11] Sharafi S, Hassanpour H. (2022). Comparison of Machine Learning Algorithms for Diabetes Prediction using Genomic Data. Journal of Medical Systems, 46(10).

[12] Sari Y, Subianto M, Rahman M. (2023). Analisis Komparatif Algoritma Klasifikasi Machine Learning untuk Prediksi Diabetes: Studi Kasus Data Klinis dan Genetik. Jurnal Teknologi Informasi dan Komunikasi, 12(1).

[13] Purnamawati A, Nugroho W, Putri D, Hidayat WF. (2020). Deteksi Penyakit Daun pada Tanaman Padi Menggunakan Algoritma Decision Tree, Random Forest, Naïve Bayes, SVM dan KNN. InfoTekJar J. Nas. Inform. dan Teknol. Jar., 5(1), 212-215.

[14] Rani, K. J. (2020). Diabetes prediction using machine learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *6*(4), 294-305.

[15] Azhari MF, Fajriyah R. (2024). Idektifikasi Gen Marker Pbmcs Ischemic Stroke Menggunakan Analisis Bioinformatika dan Support Vector Machine. Jurnal TIMES, 13(1), 73-81.

[16] Pradhan A, Sahu C. (2021). A Systematic Review of Machine Learning Techniques for Diabetes Prediction. Journal of Healthcare Engineering, 2021. (Hipotesis: SCOPUS)

[17] Dewi T, Setiawan H. (2023). Peningkatan Akurasi Prediksi Diabetes Menggunakan Algoritma Random Forest pada Data Genomik dengan Seleksi Fitur Gini Importance. Jurnal Rekayasa Informasi, 15(2).

[18] Fernández-Delgado M, Cernadas E, Barro S, Amorim D. (2021). Addressing the Class Imbalance Problem with SMOTE-Based Approaches: A Comparative Study. Expert Systems with Applications, 175.

[19] Johnson J M, Khoshgoftaar T M. (2020). Survey on the use of SMOTE for Class Imbalance Learning. Journal of Big Data, 7(1).

[20] Rasyid HA, Handayani P. (2022). Optimasi Teknik SMOTE untuk Peningkatan Performa Klasifikasi Penyakit Jantung pada Dataset Tidak Seimbang. Jurnal Informatika, 16(3).

[21] Lestari D, Wibowo A. (2023). Pengaruh Berbagai Varian SMOTE (ADASYN, Borderline) terhadap Akurasi Prediksi Diabetes Mellitus Tipe 2. Jurnal Komputasi Terapan, 7(1).

[22] Wang Q, Zhang G, Liu Y. (2024). Comprehensive Evaluation of Sampling Techniques for Imbalanced Medical Datasets. IEEE Journal of Biomedical and Health Informatics, 28(2).

[23] Utami YP, Triayudi A, Handayani EE. (2021). Sistem pakar deteksi penyakit diabetes mellitus (dm) menggunakan metode forward chaining dan certainty factor berbasis android. Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi), 5(1), 49-55.

[24] Mojrian, S., Pinter, G., Joloudari, J. H., Felde, I., Szabo-Gali, A., Nadai, L., & Mosavi, A. (2020, October). Hybrid machine learning model of extreme learning machine radial basis function for breast cancer detection and diagnosis; a multilayer fuzzy expert system. In *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)* (pp. 1-7). IEEE.

[25] Nawangnugraeni DA. (2021). Sistem pakar berbasis android untuk diagnosis diabetes melitus dengan metode forward chaining. Komputika: Jurnal Sistem Komputer, 10(1), 19-27.

[26] Sirait DA, Sitohang S. (2023). Perancangan Sistem Pakar dengan Metode Forward Chaining untuk Mendiagnosis Penyakit Diabetes Berbasis Web. Computer and Science Industrial Engineering (COMASIE), 9(2).

[27] Syahri R, Gusmaliza D, Masdalipa D. (2021). Sistem Pakar Diagnosis Penyakit Ayam Broiler Berbasis Web. Jurnal Pengembangan Sistem Informasi dan Informatika, 2.

[28] Wulandari S, Kridalaksana AH, Khairina DM. (2020). Sistem Pakar Penerapan Menu Gizi Pada Penderita Jantung Koroner Menggunakan Metode Teorema Bayes. Informatika Mulawarman: Jurnal Ilmiah Ilmu Komputer, 15(1), 1-7.

[29] Agustina N, Purwanto M. (2024). Integrating Machine Learning Prediction with Rule-Based Expert System for Clinical Decision Support in Primary Healthcare. International Journal of Medical Informatics, 185.

[30] Almeida F. (2018). Strategies to perform a mixed methods study. European Journal of Education Studies.

[31] Nayyar, A., Gadhavi, L., & Zaman, N. (2021). Machine learning in healthcare: review, opportunities and challenges. *Machine learning and the internet of medical things in healthcare*, 23-45.

[32] Letunic I, Khedkar S, Bork P. (2021). SMART: recent updates, new developments and status in 2020. Nucleic acids research, 49(D1), D458-D460.

[33] Habehh, H., & Gohel, S. (2021). Machine learning in healthcare. *Current genomics*, *22*(4), 291-300.

[34] Smith J, Williams K. (2022). Ethical Considerations in Deploying AI for Genomic Health Prediction. Journal of Medical Ethics, 48(4)

[35] Handayani S, Prasetio B. (2023). Tantangan dan Strategi Adopsi Teknologi Digital di Puskesmas Wilayah 3T Indonesia. Jurnal Kesehatan Masyarakat Nasional, 17(1).

[36] Wibowo E, Santoso A. (2024). Studi Komparatif Metode Imputasi Data Hilang pada Dataset Kesehatan. Jurnal Statistika Terapan, 5(2).

[37] Mulyadi R, Sari R. (2021). Implementasi Teknik Pseudonimisasi Data Pasien untuk Kepatuhan Etika Penelitian Kesehatan. Jurnal Keamanan Siber dan Forensik, 6(3).

[38] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning **classifiers**. *IEEE Access*, *8*, 76516-76531.