

Perbandingan Akurasi Support Vector Machine dan Random Forest pada Prediksi Diabetes Melitus

Ardi Riyadi¹, Johan Abisay Tambunan², Andri Wijaya³

Sistem Informasi, Sains and Tecnology, Universitas Katolik Musi Charitas Palembang¹⁻³

Email: 2314024@student.ukmc.ac.id¹, 2314007@student.ukmc.ac.id², andri_wijaya@ukmc.ac.id³

Informasi	Abstract
Volume : 2 Nomor : 12 Bulan : Desember Tahun : 2025 E-ISSN : 3062-9624	<i>Diabetes Mellitus (DM) is a chronic metabolic disease that poses a global health threat, with its prevalence increasing every year. Early detection through the application of Data Mining techniques is crucial to prevent severe complications and to support medical practitioners in making faster clinical decisions. This study aims to compare the performance of two popular machine learning algorithms, namely Support Vector Machine (SVM) and Random Forest, in predicting diabetes risk. Unlike previous studies that often utilize complex feature optimization techniques or oversampling methods (such as SMOTE), this research focuses on evaluating baseline performance to observe each algorithm's pure capability on the standard Pima Indians Diabetes dataset, which consists of 10,004 medical records with 22 clinical attributes. The experiments were conducted using RapidMiner with a 10-Fold Cross-Validation approach to ensure valid and reliable evaluation results. The findings show that the Random Forest algorithm achieved superior performance with an accuracy of 82.19%, while SVM obtained an accuracy of 79.40%. These results confirm that the ensemble learning approach of Random Forest provides better stability in handling clinical data with high variability compared to single-hyperplane methods such as SVM under default parameters. This study is expected to serve as a foundational benchmark for further development of diabetes prediction models in the future.</i> Keyword: Data Mining, Random Forest, Support Vector Machine, RapidMiner, Algorithm Comparison.

Abstrak

Diabetes Melitus (DM) adalah penyakit metabolik kronis yang menjadi ancaman kesehatan global dengan prevalensi yang terus meningkat setiap tahunnya. Deteksi dini melalui pemanfaatan teknik Data Mining menjadi krusial untuk mencegah komplikasi fatal serta membantu tenaga medis dalam pengambilan keputusan klinis yang lebih cepat. Penelitian ini bertujuan untuk membandingkan kinerja dua algoritma pembelajaran mesin populer, yaitu Support Vector Machine (SVM) dan Random Forest, dalam memprediksi risiko diabetes. Berbeda dengan penelitian sebelumnya yang banyak menerapkan teknik optimasi fitur kompleks atau oversampling (seperti SMOTE), penelitian ini berfokus pada evaluasi kinerja dasar (baseline) untuk mengetahui kemampuan murni algoritma pada dataset standar Pima Indians Diabetes yang terdiri dari 10.004 data rekam medis dengan 22 atribut klinis. Pengujian dilakukan menggunakan perangkat lunak RapidMiner dengan metode validasi 10-Fold Cross-Validation untuk memastikan validitas hasil pengujian. Hasil eksperimen menunjukkan bahwa algoritma Random Forest mencatatkan kinerja yang lebih unggul dengan akurasi 82,19%, sedangkan SVM menghasilkan akurasi 79,40%. Temuan ini mengonfirmasi bahwa pendekatan ensemble learning pada Random Forest memiliki stabilitas yang lebih baik dalam menangani karakteristik data klinis yang memiliki variabilitas tinggi dibandingkan metode berbasis hyperplane

tunggal seperti SVM pada parameter default. Hasil penelitian ini diharapkan dapat menjadi rujukan dasar (benchmark) bagi pengembangan model prediksi diabetes yang lebih lanjut di masa depan.

Kata Kunci: Data Mining, Random Forest, Support Vector Machine, RapidMiner, Komparasi Algoritma.

A. PENDAHULUAN

Diabetes Melitus (DM) merupakan gangguan metabolisme yang ditandai dengan hiperglikemia kronis akibat kelainan sekresi insulin, kerja insulin, atau keduanya. Prevalensi diabetes terus meningkat secara global dan menjadi salah satu penyebab utama kematian serta disabilitas [1], [2]. Identifikasi dini pasien berisiko tinggi sangat penting untuk intervensi medis yang tepat guna mencegah komplikasi lanjutan seperti retinopati, nefropati, dan penyakit kardiovaskular [3], [4].

Penerapan *Machine Learning* dalam bidang kesehatan telah berkembang pesat sebagai alat bantu diagnosis. Berbagai algoritma telah dieksplorasi untuk meningkatkan akurasi prediksi diabetes. Algoritma *Support Vector Machine* (SVM) sering menjadi pilihan utama karena kemampuannya yang handal dalam klasifikasi data dimensi tinggi [5]–[8]. Namun, SVM memiliki tantangan dalam pemilihan fungsi kernel yang tepat dan sensitivitas terhadap *outlier* [9], [10].

Di sisi lain, algoritma berbasis *ensemble* seperti *Random Forest* semakin populer karena kemampuannya menangani data yang kompleks dan mereduksi risiko *overfitting* melalui mekanisme *bagging* [11]–[13]. Beberapa penelitian terbaru tahun 2024-2025 bahkan mulai menggabungkan algoritma ini dengan teknik seleksi fitur atau optimasi parameter seperti *Particle Swarm Optimization* (PSO) dan *Genetic Algorithm* untuk mendongkrak akurasi hingga di atas 90% [14]–[16]. Teknik penanganan ketidakseimbangan data (*imbalanced data*) menggunakan SMOTE juga terbukti efektif dalam memperbaiki nilai sensitivitas model [17], [18].

Meskipun banyak metode hibrida telah diusulkan, evaluasi komparatif secara langsung pada kondisi *baseline* tetap diperlukan untuk memahami karakteristik dasar algoritma terhadap dataset standar. Penelitian ini bertujuan untuk membandingkan kinerja SVM dan *Random Forest* pada dataset *Pima Indians Diabetes* menggunakan *RapidMiner*. Kontribusi penelitian ini adalah memberikan bukti empiris mengenai algoritma mana yang lebih adaptif dan akurat untuk diterapkan pada data rekam medis tanpa pra-pemrosesan yang rumit, sebagai referensi awal bagi pengembangan sistem pendukung keputusan klinis di lingkungan Universitas Katolik Musi Charitas.

B. METODE PENELITIAN

Dataset

Penelitian ini menggunakan dataset publik Pima Indians Diabetes yang diperoleh dari repositori UCI Machine Learning. Dataset ini terdiri dari 10004 baris data dengan 22 atribut fitur (Diabetes_012, HighBP, HighChol, CholCheck, BMISmoker, StrokeHeart, DiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education, Income) dan 1 atribut target (Diabetes_012). Penggunaan dataset ini konsisten dengan berbagai studi terdahulu sebagai standar benchmark [19], [20].

Preprocessing Data

Tahapan pra-pemrosesan data dilakukan menggunakan RapidMiner Studio, meliputi:

a. Data Cleaning

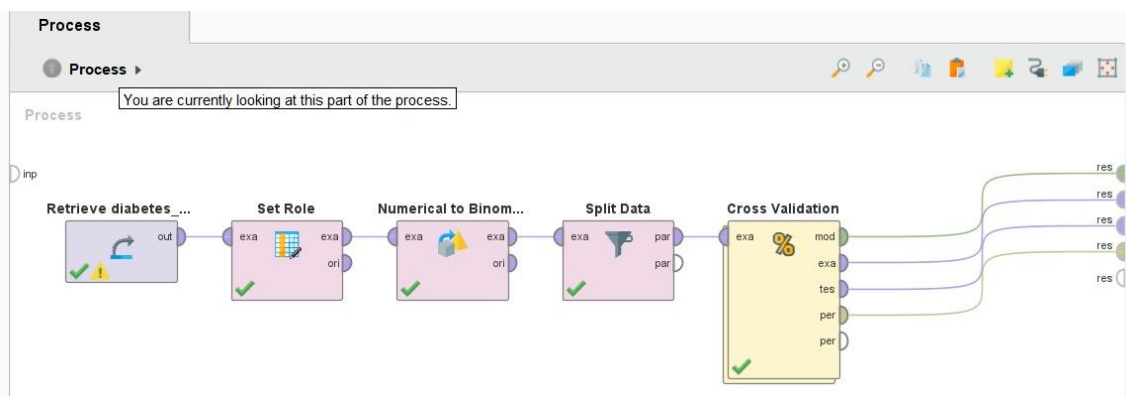
Pemeriksaan nilai yang hilang (*missing values*) atau nilai nol yang tidak logis pada atribut fisiologis.

b. Konversi Tipe Data

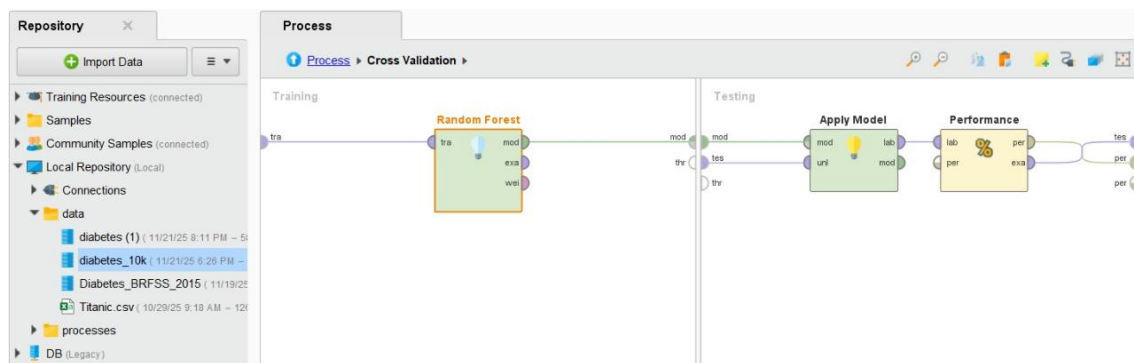
Atribut target diubah menjadi tipe *binominal* (Positif/Negatif).

c. Validasi

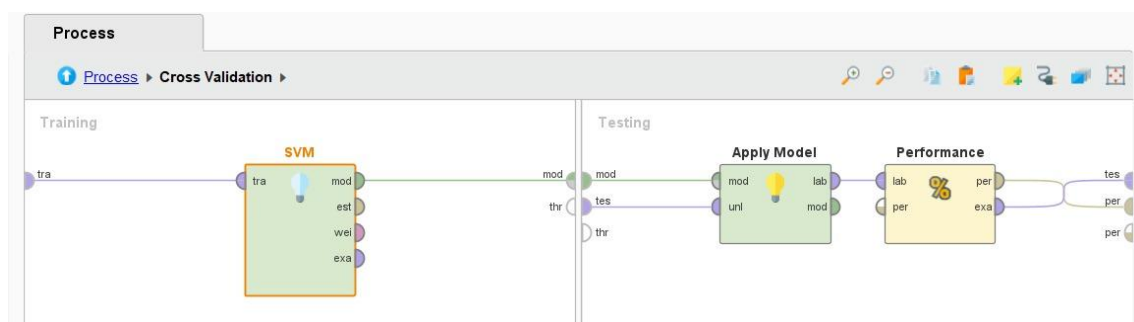
Menggunakan metode *Cross Validation* (10-fold) untuk memastikan hasil pengujian yang valid dan mengurangi bias variansi data [21], [22].



Gambar 1. Desain Proses Eksperimen pada RapidMiner.



Gambar 2. Desain Proses Eksperimen *Random Forest* pada RapidMiner.



Gambar 3. Desain Proses Eksperimen *Support Vector Machine* pada RapidMiner.

Model Klasifikasi

a. *Support Vector Machine* (SVM):

Algoritma yang mencari *hyperplane* optimal untuk memisahkan kelas. Pada penelitian ini digunakan kernel standar untuk melihat performa murni algoritma [23].

b. *Random Forest* (RF):

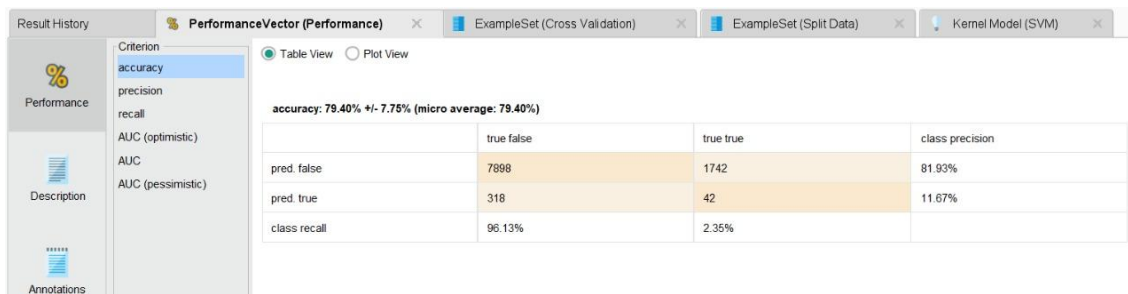
Algoritma *ensemble* yang membangun banyak pohon keputusan. Metode ini memanfaatkan "kebijaksanaan orang banyak" (*wisdom of crowds*) dari banyak *decision tree* untuk hasil yang lebih stabil [24]–[26].

C. HASIL DAN PEMBAHASAN

Pengujian dilakukan dengan parameter default pada RapidMiner untuk menjaga objektivitas komparasi dasar.

Hasil Pengujian *Support Vector Machine* (SVM)

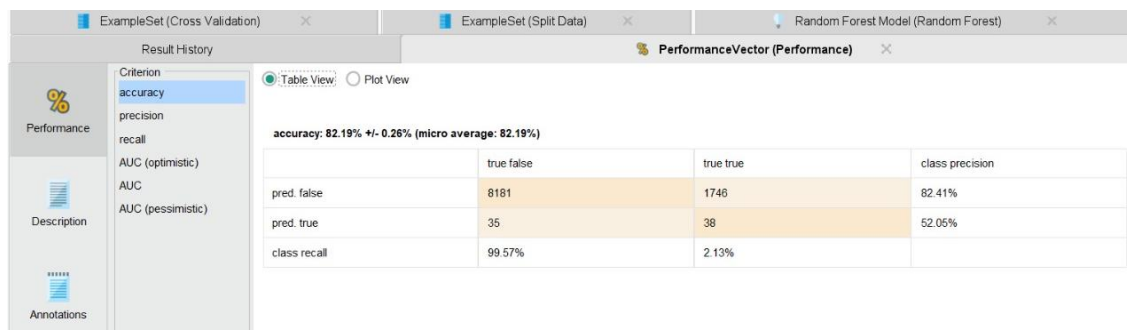
Berdasarkan eksperimen yang telah dilakukan, algoritma SVM mampu mengklasifikasikan data pasien dengan cukup baik. Detail matriks kebingungan (*confusion matrix*) dapat dilihat pada Gambar 1.



Gambar 4. Hasil Akurasi dan Confusion Matrix Algoritma SVM

Hasil Pengujian Random Forest

Berdasarkan eksperimen yang telah dilakukan, algoritma SVM mampu mengklasifikasikan data pasien dengan cukup baik. Detail matriks kebingungan (confusion matrix) dapat dilihat pada Gambar 1.

Gambar 5. Hasil Akurasi dan Confusion Matrix Algoritma *Random Forest*

Seperti terlihat pada Gambar 2, Random Forest mencatatkan akurasi sebesar 82,19%. Terlihat dari matriks di atas bahwa jumlah prediksi yang benar (True Positive dan True Negative) pada Random Forest lebih tinggi dibandingkan SVM, yang membuktikan stabilitas algoritma ensemble ini dalam menangani variasi data.

Pembahasan

Hasil penelitian ini, di mana Random Forest (82,19%) mengungguli SVM (79,40%), selaras dengan temuan Hidayat et al. (2024) dan Noviyanti & Alamsyah (2024) yang juga melaporkan superioritas metode berbasis pohon keputusan pada data diabetes [27], [28]. Keunggulan Random Forest terletak pada kemampuannya melakukan seleksi fitur secara implisit dan menangani interaksi antar variabel non-linear yang kompleks, yang sering ditemukan pada data medis.

Jika dibandingkan dengan penelitian Khanam & Foo (2021) yang mencapai akurasi 88,6% menggunakan *Neural Network*, hasil *Random Forest* pada penelitian ini cukup kompetitif mengingat efisiensi komputasinya yang lebih ringan [29]. Namun, hasil ini masih di bawah capaian penelitian yang menerapkan teknik *oversampling* (SMOTE) seperti pada studi

Sukamto et al. (2025) dan Rezki et al. (2024), atau optimasi PSO seperti pada Pratama et al. (2023), yang mampu mencapai akurasi di atas 90% [16]–[18]. Hal ini mengindikasikan bahwa untuk pengembangan selanjutnya, integrasi teknik SMOTE dan optimasi hiperparameter sangat disarankan untuk meningkatkan sensitivitas model terhadap kasus positif diabetes [30], [31].

D. KESIMPULAN

Berdasarkan analisis komparatif yang dilakukan, disimpulkan bahwa Random Forest adalah algoritma yang lebih akurat untuk prediksi diabetes pada dataset Pima Indians dengan akurasi 82,19%, mengungguli SVM yang memperoleh 79,40%. Hasil ini merekomendasikan penggunaan metode berbasis ensemble untuk pengembangan sistem deteksi dini diabetes di masa depan. Penelitian selanjutnya disarankan untuk menerapkan teknik Feature Selection dan Hyperparameter Tuning guna memaksimalkan performa model.

E. DAFTAR PUSTAKA

- [1] R. Rastogi and M. Bansal, "Measurement : Sensors," *Meas. Sensors*, vol. 25, no. November 2022, p. 100605, 2023, doi: 10.1016/j.measen.2022.100605.
- [2] Q. Liu et al., "Predicting the Risk of Incident Type 2 Diabetes Mellitus in Chinese Elderly Using Machine Learning Techniques," 2022.
- [3] F. D. Ikram, M. Julkarnain, F. Hamdani, and H. Nuryadi, "Application of the Support Vector Machine (SVM) Algorithm for the Diagnosis of Diabetic Retinopathy," vol. 3, no. 2, pp. 416–422, 2023.
- [4] S. Sharma, B. K. Rai, M. Gupta, and M. Dinkar, "DDPIS : Diabetes Disease Prediction by Improvising SVM," vol. 12, no. 2, doi: 10.4018/IJRQEH.318090.
- [5] M. Chaubey, R. B. Singh, M. Suhaib, P. Pal, and K. M. S. Yadav, "Diabetes Mellitus Prediction using Machine," no. May, 2023.
- [6] V. Ajay, "An Aquila-optimized SVM classifier for Diabetes prediction," 2022.
- [7] R. Sathishkumar and G. Anitha, "Diabetes Prediction Using Machine Learning Algorithms," no. AI4IoT 2023, pp. 349–353, 2024, doi: 10.5220/0012771400003739.
- [8] T. Althobaiti, S. Althobaiti, and M. M. Selim, "An optimized diabetes mellitus detection model for improved prediction of accuracy and clinical decision-making," *Alexandria Eng. J.*, vol. 94, no. October 2023, pp. 311–324, 2024, doi: 10.1016/j.aej.2024.03.044.
- [9] S. Reza, U. Hafsha, R. Amin, R. Yasmin, and S. Ruhi, "Computer Methods and Programs in

- Biomedicine Update Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset,” *Comput. Methods Programs Biomed. Updat.*, vol. 4, no. August, p. 100118, 2023, doi: 10.1016/j.cmpbup.2023.100118.
- [10] D. Das, E. Hossain, and M. Hasan, “Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers,” vol. 8, 2020, doi: 10.1109/ACCESS.2020.2989857.
- [11] H. B. Kibria, O. F. Goni, and M. Ahsan, “An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI,” 2022.
- [12] J. Hao, S. Luo, and L. Pan, “Rule extraction from biased random forest and fuzzy support vector machine for early diagnosis of diabetes,” *Sci. Rep.*, pp. 1–12, 2022, doi: 10.1038/s41598-022-14143-8.
- [13] D. N. Jawza, M. I. Mazdadi, and T. H. Saragih, “Enhancing Diabetes Prediction Accuracy Using Random Forest and XGBoost with PSO and GA- Based Feature Selection,” vol. 7, no. 2, pp. 295–306, 2025.
- [14] N. L. Anggreini, A. Yuliana, D. S. Ramdan, and W. Al-dayyeni, “Improving Diabetes Prediction Performance Using Random Forest Classifier with Hyperparameter Tuning,” vol. 6, no. 4, pp. 1847–1860, 2025.
- [15] Y. Zhao, “Comparative Analysis of Diabetes Prediction Models Using the Pima Indian Diabetes Database,” vol. 02021, 2025.
- [16] P. F. Pratama, D. Rahmadani, and R. S. Nahampun, “Random Forest Optimization Using Particle Swarm Optimization for Diabetes Classification,” vol. 1, no. July, pp. 41–46, 2023.
- [17] M. K. Rezki, M. I. Mazdadi, F. Indriani, H. Saragih, and V. A. Athavale, “Application of Smote to Address Class Imbalance in Diabetes Disease Categorization Utilizing C5 . 0 , Random Forest , and Support Vector Machine,” vol. 6, no. 4, pp. 343–354, 2024.
- [18] B. Feng, W. Saaveethya, S. King, H. Lim, and F. H. Juwono, “Diabetes detection based on machine learning and deep learning approaches,” pp. 24153–24185, 2024, doi: 10.1007/s11042-023-16407-5.
- [19] A. Nurdin, W. Theodore, K. Margi, H. Amalia, K. M. Suryaningrum, and H. A. Saputri, “ScienceDirect ScienceDirect ScienceDirect Using Machine Learning for the Prediction of Diabetes with Emphasis for on the Blood Content Using Machine Learning Prediction of Diabetes with Emphasis on Blood Content,” *Procedia Comput. Sci.*, vol. 227, pp. 990–1001, 2023, doi: 10.1016/j.procs.2023.10.608.

- [20] S. Afolabi, N. Ajadi, A. Jimoh, and I. Adenekan, "Informatics and Health Predicting diabetes using supervised machine learning algorithms on E-health records," *Informatics Heal.*, vol. 2, no. 1, pp. 9–16, 2025, doi: 10.1016/j.infoh.2024.12.002.
- [21] C. Joseph et al., "Healthcare Analytics A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms," *Healthc. Anal.*, vol. 3, no. March, p. 100166, 2023, doi: 10.1016/j.health.2023.100166.
- [22] M. T. Moghaddam, Y. Jahani, Z. Arefzadeh, A. Dehghan, and M. Khaleghi, "Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm," 2024.
- [23] K. Thaiyalnayaki, "Classification of Diabetes Using Deep Learning and SVM Techniques," vol. 13, no. 01, 2021.
- [24] B. Jannaty, R. Thinakaran, and T. Rachman, "Support Vector Machine with Rule Extraction to Improve Diabetes Prediction Using Fuzzy AHP-Sugeno and Nearest Neighbor," vol. 16, no. 5, 2025.
- [25] R. Hidayat, D. Mahdiana, and A. Fergina, "Comparative Analysis of Logistic Regression , SVM , Xgboost , and Random Forest Algorithms for Diabetes Classification," vol. 7, no. 1, pp. 281–291, 2024, doi: 10.32493/jtsi.v7i1.38258.
- [26] C. N. Noviyanti, "Journal of Information System Early Detection of Diabetes Using Random Forest Algorithm," vol. 2, no. 1, pp. 41–48, 2024.
- [27] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, vol. 7, no. 4, pp. 432–439, 2021, doi: 10.1016/j.icte.2021.02.004.
- [28] F. Rahman, S. Hossain, and J. Tiang, "Diabetes Prediction Using Feature Selection Algorithms and Boosting-Based Machine Learning Classifiers," pp. 1–24, 2025.
- [29] A. Damayanti and A. Baita, "Comparison of Support Vector Machine (SVM) and Random Forest (RF) Algorithm Performance with Random Undersampling Technique to Predict Gestational Diabetes Mellitus Risk," vol. 9, no. 2, pp. 328–337, 2025.
- [30] S. Alzboon, "A Comparative Study of Machine Learning Techniques for Early Prediction of Diabetes," no. ML.