202110370311159
M.Rafly Rahman
Prodi Informatika

INTEGRASI DATA TABULAR DAN REPRESENTASI TEKS UNTUK PREDIKSI RISIKO KLINIS MENGGUNAKAN MACHINE LEARNING DAN LARGE LANGUAGE MODELS

Laporan Tugas Akhir

Diajukan Untuk Memenuhi Persyaratan Guna Meraih Gelar Sarjana Informatika
Universitas Muhammadiyah Malang

M. Rafly Rahman
202110370311159

Bidang Minat
Data Science

PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS MUHAMMADIYAH MALANG
2025/2026

202110370311159
M.Rafly Rahman
Prodi Informatika

LEMBAR PERSETUJUAN

LEMBAR PERSETUJUAN

INTEGRASI DATA TABULAR DAN REPRESENTASI TEKS UNTUK PREDIKSI RISIKO KLINIS MENGGUNAKAN MACHINE LEARNING DAN LARGE LANGUAGE MODELS

TUGAS AKHIR

Sebagai Persyratan Guna Meraih Gelar Sarjana Strata 1
Informatika Universitas Muhammadiyah Malang

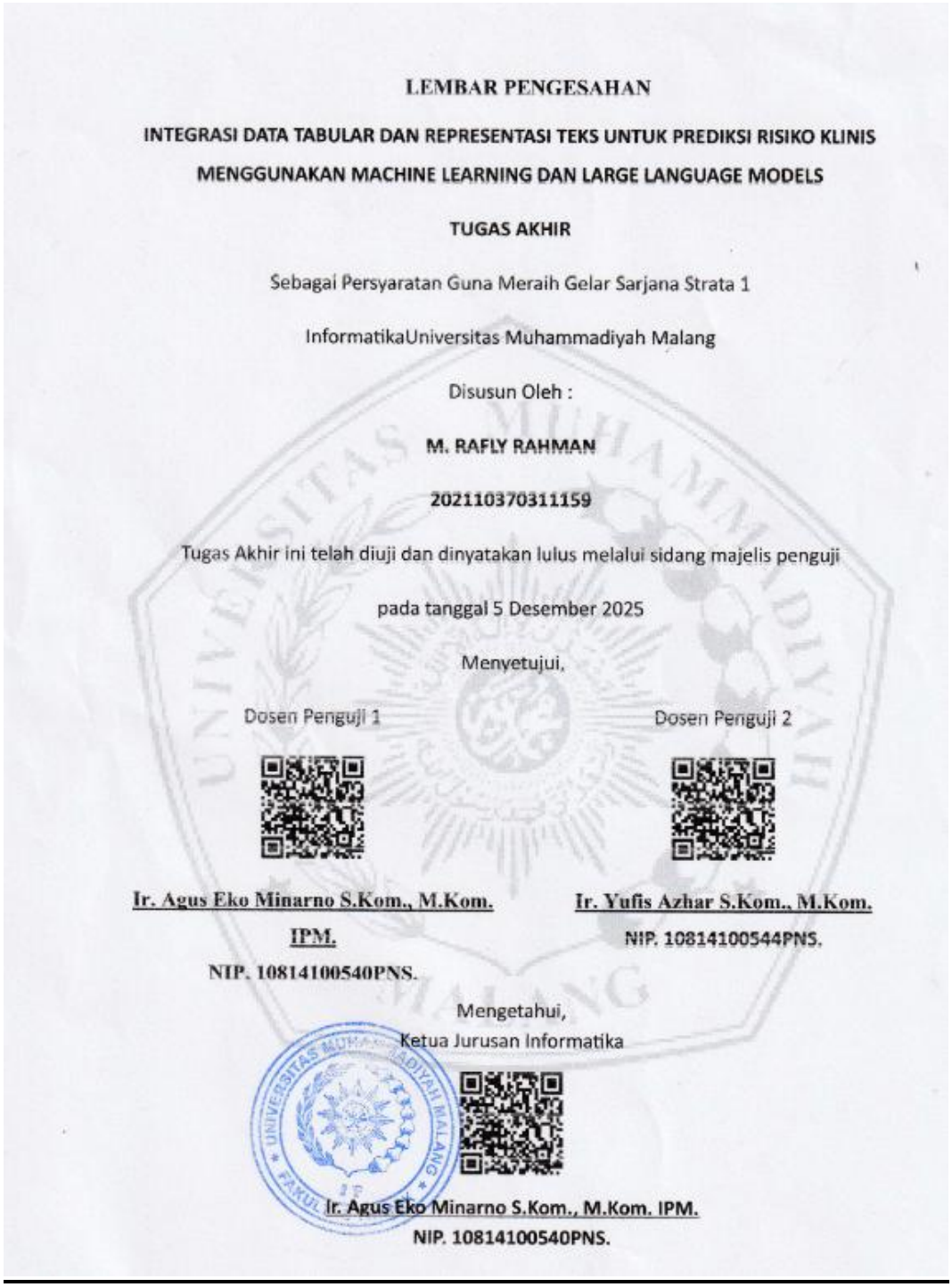
Menyetuji,
Malang, 29 September 2025

Dosen Pembimbing 1

Setio Basuki M.T., Ph.D.
NIP. 10809070477PNS

202110370311159
M.Rafly Rahman
Prodi Informatika

LEMBAR PENGESAHAN



LEMBAR PENGESAHAN

INTEGRASI DATA TABULAR DAN REPRESENTASI TEKS UNTUK PREDIKSI RISIKO KLINIS MENGGUNAKAN MACHINE LEARNING DAN LARGE LANGUAGE MODELS

TUGAS AKHIR

Sebagai Persyaratan Guna Meraih Gelar Sarjana Strata 1

InformatikaUniversitas Muhammadiyah Malang

Disusun Oleh :

M. RAFLY RAHMAN

202110370311159

Tugas Akhir ini telah diuji dan dinyatakan lulus melalui sidang majelis penguji

pada tanggal 5 Desember 2025

Menyetujui,

Dosen Penguji 1

Ir. Agus Eko Minarno S.Kom., M.Kom. IPM.

NIP. 10814100540PNS.

Dosen Penguji 2

Ir. Yufis Azhar S.Kom., M.Kom.

NIP. 10814100544PNS.

Mengetahui,

Ketua Jurusan Informatika

Ir. Agus Eko Minarno S.Kom., M.Kom. IPM.

NIP. 10814100540PNS.

202110370311159
M.Rafly Rahman
Prodi Informatika

LEMBAR PERNYATAAN

LEMBAR PERNYATAAN

Yang bertanda tangan dibawah ini:

Nama : **M. RAFLY RAHMAN**
NIM : **202110370311159**
FAK./JUR. : **INFORMATIKA**

Dengan ini saya menyatakan bahwa Tugas Akhir dengan judul **"INTEGRASI DATA TABULAR DAN REPRESENTASI TEKS UNTUK PREDIKSI RISIKO KLINIS MENGGUNAKAN MACHINE LEARNING DAN LARGE LANGUAGE MODELS"** beserta seluruh isinya adalah karya saya sendiri dan bukan merupakan karya tulis orang lain, baik sebagian maupun seluruhnya, kecuali kutipan yang telah disebutkan sumbernya.

Demikian surat pernyataan ini saya buat dengan sebenar-benarnya. Apabila kemudian ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya ini, atau ada klaim dari pihak lain terhadap keaslian karya saya ini maka saya siap menanggung segala bentuk resiko/sanksi yang berlaku.

Mengetahui,
Dosen Pembimbing

Setio Basuki M.T., Ph.D.

Malang, 29 September 2025
Yang Membuat Pernyataan

M. Rafly Rahman

ii

202110370311159
M.Rafly Rahman
Prodi Informatika

ABSTRAK

Kesehatan global saat ini menghadapi tantangan serius akibat meningkatnya jumlah pasien dengan penyakit kronis, seperti gagal jantung, diabetes, dan kanker. Masalah ini muncul dari keterbatasan sistem rekam medis elektronik (EHR), yang belum sepenuhnya mampu menjamin ketepatan diagnosis klinis karena kemungkinan kesalahan input data dan keterlambatan dalam identifikasi gejala oleh tenaga medis. Menanggapi masalah tersebut, makalah ini fokus pada integrasi data tabular medis dengan pendekatan klasifikasi berbasis classical machine learning (ML) dan large language models (LLM) untuk meningkatkan akurasi prediksi diagnosis pasien. Makalah ini bertujuan untuk mengembangkan dan membandingkan kinerja berbagai model ML, seperti XGBoost, SVM, dan Logistic Regression, serta model LLM seperti Gemini, LLaMA, dan Qwen dalam skenario fine-tuning, few-shot, dan zero-shot. Hasil penelitian menunjukkan bahwa kombinasi Llama dengan pendekatan few-shot (250 shots) mencapai akurasi tertinggi hingga 96,0% dalam memprediksi risiko gagal jantung. Temuan utama dari penelitian ini adalah bahwa representasi teks naratif dari data tabular yang diproses dengan LLM secara signifikan meningkatkan pemahaman kontekstual dan akurasi klasifikasi, sehingga pendekatan ini sangat potensial untuk diterapkan dalam pengambilan keputusan klinis berbasis AI.

Kata Kunci: Medical Data Data tabular, *Large Language Models* (LLM), *Clinical Risk Prediction*, Serialisasi, *Few-shot Learning*

202110370311159
M.Rafly Rahman
Prodi Informatika

ABSTRACT

Global health is currently facing serious challenges due to the increasing number of chronic disease patients such as heart failure, diabetes, and cancer. This issue arises from the limitations of electronic health record (EHR) systems, which are not yet fully capable of ensuring accurate clinical diagnoses because of potential data input errors and delays in symptom identification by medical personnel. In response to this issue, this paper focuses on the integration of medical tabular data with a classification approach based on classical machine learning (ML) and large language models (LLM) to improve the accuracy of patient diagnosis predictions. This paper aims to develop and compare the performance of various ML models, such as XGBoost, SVM, and Logistic Regression, as well as LLM models like Gemini, LLaMA, and Qwen in fine-tuning, few-shot, and zero-shot scenarios. The paper results show that the combination of Llama and the few-shot approach (250 shots) achieved the highest accuracy of up to 96.0%, in predicting heart failure risk. The main finding of this study is that the narrative text representation of tabular data processed with LLM significantly enhances contextual understanding and classification accuracy, making this approach highly potent for application in AI-based clinical decision-making.

Keywords: Medical Tabular Data, Large Language Models (LLM), Clinical Risk Prediction, Data Serialization, Few-shot Learning

202110370311159
M.Rafly Rahman
Prodi Informatika

LEMBAR PERSEMBAHAN

Puji syukur kehadirat Allah SWT atas segala rahmat, karunia, dan hidayah-Nya sehingga penulis dapat menyelesaikan Tugas Akhir ini dengan baik. Penyusunan laporan ini tidak lepas dari bantuan, bimbingan, serta dukungan dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Allah SWT, yang senantiasa memberikan kehidupan, kesehatan, kekuatan, serta kemudahan dalam setiap langkah penulis.
2. Kedua orang tua tercinta, Bapak Dicky Zulkarnain dan Ibu Enny Fitriana, yang selalu memberikan kasih sayang, doa, semangat, dan dukungan penuh sehingga penulis dapat menyelesaikan studi ini.
3. Bapak Setio Basuki, MT., Ph.D., selaku dosen pembimbing tugas akhir, atas segala waktu, bimbingan, arahan, dan kesabaran dalam membantu penulis menyelesaikan penelitian ini.
4. Bapak/Ibu Dekan Fakultas Teknik Universitas Muhammadiyah Malang, atas segala fasilitas dan dukungan yang diberikan.
5. Bapak Dr. Ir. Agus Eko Minarno, S.Kom., M.Kom, selaku Ketua Jurusan Teknik Informatika Universitas Muhammadiyah Malang, atas ilmu, arahan, dan kesempatan yang telah diberikan.
6. Seluruh dosen dan staf Program Studi Informatika Universitas Muhammadiyah Malang, atas ilmu, pengalaman, dan bimbingan yang diberikan selama masa perkuliahan.
7. Dinda Agustina, yang telah memberikan dukungan, semangat, dan saran selama proses perkuliahan hingga penyusunan Tugas Akhir ini.
8. Diri penulis sendiri, atas usaha, doa, dan semangat yang telah diberikan untuk menyelesaikan Tugas Akhir ini hingga tuntas.

202110370311159
M.Rafly Rahman
Prodi Informatika

KATA PENGANTAR

Puji syukur peneliti panjatkan ke hadirat Allah SWT atas limpahan rahmat, taufik, dan hidayah-Nya sehingga peneliti dapat menyelesaikan tugas akhir yang berjudul:

"INTEGRASI DATA TABULAR DAN REPRESENTASI TEKS UNTUK PREDIKSI RISIKO KLINIS MENGGUNAKAN MACHINE LEARNING DAN LARGE LANGUAGE MODELS"

Tugas akhir ini disusun sebagai salah satu syarat untuk menyelesaikan studi dan memperoleh gelar sarjana. Di dalamnya disajikan pokok-pokok pembahasan yang mencakup latar belakang penelitian, metode yang digunakan, serta hasil dan pembahasan yang disimpulkan berdasarkan proses penelitian yang telah dilakukan.

Peneliti menyadari sepenuhnya bahwa dalam penyusunan tugas akhir ini masih terdapat berbagai kekurangan dan keterbatasan. Oleh karena itu, peneliti dengan rendah hati mengharapkan kritik dan saran yang membangun dari semua pihak demi kesempurnaan penulisan ini. Besar harapan peneliti, semoga tugas akhir ini dapat memberikan manfaat serta menjadi salah satu kontribusi kecil dalam pengembangan ilmu pengetahuan, khususnya di bidang teknologi dan analisis data.

202110370311159
M.Rafly Rahman
Prodi Informatika

DAFTAR ISI

LEMBAR PERSETUJUAN ..... **i**
LEMBAR PENGESAHAN ..... **ii**
LEMBAR PERNYATAAN ..... **iii**
ABSTRAK ..... **iv**
ABSTRACT ..... **v**
LEMBAR PERSEMBAHAN ..... **vi**
KATA PENGANTAR ..... **vii**
DAFTAR TABEL ..... **xi**
DAFTAR GAMBAR ..... **xii**
BAB I PENDAHULUAN **1**
1.1 Latar Belakang Masalah 1
1.2 Rumusan Masalah 3
1.3 Tujuan Penelitian 4
1.4 Batasan Masalah 4
BAB II TINJAUAN PUSTAKA **6**
2.1 Studi Literatur 6
2.2 Gagal Jantung 8
2.3 Diabetes 8
2.4 Kanker 9
2.5 Data Tabular 9
2.6 Machine Learning 9
2.6.1 Logistic Regression 10
2.6.2 Support Vector Machine (SVM) 10
2.6.3 Naive Bayes 10
2.6.4 XGBoost 11
2.7 Serialisasi 11
2.8 Teks Repsentasi 11
2.8.1. Term Frequency–Inverse Document Frequency (TF IDF) 12
2.8.2. Bag of Word 12

2.8.3. Word2Vec 12
2.8.4. N-Grams 13
2.9 Large Language Model 13
2.9.1 Prompt Engineering 14
2.9.2 Few-Shot 14
2.9.3 Zero-Shot 14
2.9.4 Fine Tuning 15
2.9.5 Large Language Model Meta AI (LLaMA) 15
2.9.6 Gemini 15
2.9.7 Qwen 16
2.10 Evalusiasi Model 16
BAB III METODOLOGI PENELITIAN 18
3.1 Tahapan Penelitian 18
3.2 Dataset 18
3.3 Klasifikasi pada Dataset Tabular Original 20
3.4 Klasifikasi pada Dataset yang Diserialisasi menggunakan Metode NLP Klasik 20
3.5 Klasifikasi pada Dataset yang Diserialisasi menggunakan Large Language Model 22
3.5.1 Skenario Pembelajaran Fine-Tuning untuk Klasifikasi 23
3.5.2 Skenario Pembelajaran Few-shot untuk Klasifikasi 24
3.5.3 Skenario Pembelajaran Zero-Shot Untuk Klasifikasi 24
BAB IV HASIL DAN PEMBAHASAN 25
4.1 Hasil Klasifikasi pada Dataset Tabular Original 25
4.2 Hasil Klasifikasi pada Dataset Serialisasi menggunakan Metode NLP Klasik 27
4.3 Hasil Klasifikasi pada Dataset Serialisasi Menggunakan Large Langauge Model: Fine-Tuning 29
4.4 Hasil Klasifikasi pada Dataset Serialisasi Menggunakan Large Language Model: Zero-shot and Few-shot 30
4.5 Analisis Perbandingan Hasil Klasifikasi Model Terbaik 31
4.6 Analisis Perbandingan Hasil Klasifikasi Pada Penelitian Terdahulu 33

BAB V KESIMPULAN DAN SARAN 35
5.1 Kesimpulan 35
5.2 Saran..... 35
DAFTAR PUSTAKA..... 37

202110370311159
M.Rafly Rahman
Prodi Informatika

DAFTAR TABEL

Tabel 2. 1. Penelitian Terdahulu 6
Tabel 3. 1. Dataset Diabetes 19
Tabel 3. 2. Dataset Prediksi Kanker 19
Tabel 3. 3. Dataset Gagal Jantung 19
Tabel 4. 1. Persentase Kinerja pada Dataset Tabular Original 27
Tabel 4. 2. Persentase Performa pada Dataset Serialisasi Menggunakan Metode NLP 28
Tabel 4. 3. Performa Model pada Skenario Fine-Turning LLM 30
Tabel 4. 4. Akurasi LLM dalam Skenario Beberapa Tembakan dan Tanpa Tembakan 31
Tabel 4. 5. Perbandingan Akurasi Model Klasifikasi Medis pada Pendekatan Tabular Original, NLP, dan LLM 32
Tabel 4. 6. Perbandingan Hasil Klasifikasi 34

202110370311159
M.Rafly Rahman
Prodi Informatika

DAFTAR GAMBAR

Gambar 3. 1. Kerangka Klasifikasi Dataset Tabular menggunakan ML dan LLM 18
Gambar 3. 2. Hasil Serialisasi pada Dataset Tabular Asli 22

202110370311159
M.Rafly Rahman
Prodi Informatika

UNIVERSITAS MUHAMMADIYAH MALANG

FAKULTAS TEKNIK

INFORMATIKA
informatika.umm.ac.id | informatika@umm.ac.id

FORM CEK PLAGIARISME LAPORAN TUGAS AKHIR

Nama Mahasiswa : M RAFLY RAHMAN

NIM : 202110370311159

Judul TA : INTEGRASI DATA TABULAR DAN REPRESENTASI TEKS UNTUK PREDIKSI RISIKO KLINIS MENGGUNAKAN MACHINE LEARNING DAN LARGE LANGUAGE MODELS

Hasil Cek Plagiarisme dengan Turnitin

No.	Komponen Pengecekan	Nilai Maksimal Plagiarisme (%)	Hasil Cek Plagiarisme (%) *
1.	Bab 1 – Pendahuluan	10 %	7 %
2.	Bab 2 – Daftar Pustaka	25 %	3 %
3.	Bab 3 – Analisis dan Perancangan	25 %	2 %
4.	Bab 4 – Implementasi dan Pengujian	15 %	2 %
5.	Bab 5 – Kesimpulan dan Saran	5 %	3 %
6.	Makalah Tugas Akhir	20%	0 %

**) Hasil cek plagiarism diisi oleh pemeriksa (staf TU)*

***) Maksimal 5 kali (4 Kali sebelum ujian, 1 kali sesudah ujian)**

Mengetahui,

Pemeriksa (Staff TU)

(.....)

DAFTAR PUSTAKA

[1] Khariri and L. Andriani, "Dominasi Penyakit Tidak Menular dan Pola Makan Yang Tidak Sehat," *Pros Sem Nas Masy Biodiv Indon*, vol. 6, no. 1, pp. 649–652, 2020, doi: 10.13057/psnmbi/m060127.

[2] W. H. O. WHO, "Noncommunicable diseases," *World Health Organization (WHO)*, 2024. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

[3] T. N. Bogale *et al.*, "Effect of electronic records on mortality among patients in hospital and primary healthcare settings: a systematic review and meta-analyses," *Front. Digit. Heal.*, vol. 6, no. June, pp. 1–14, 2024, doi: 10.3389/fdgth.2024.1377826.

[4] T. S. Hwang, M. Thomas, M. Hribar, A. Chen, and E. White, "The Impact of Documentation Workflow on the Accuracy of the Coded Diagnoses in the Electronic Health Record," *Ophthalmol. Sci.*, vol. 4, no. 1, p. 100409, 2024, doi: 10.1016/j.xops.2023.100409.

[5] R. A. Dixit, C. L. Boxley, S. Samuel, V. Mohan, R. M. Ratwani, and J. A. Gold, "Electronic Health Record Use Issues and Diagnostic Error: A Scoping Review and Framework," *J. Patient Saf.*, vol. 19, no. 1, pp. E25–E30, 2023, doi: 10.1097/PTS.0000000000001081.

[6] R. M. Ratwani, D. W. Bates, and J. Gold, "Addressing Electronic Health Record Contributions To Diagnostic Error," *Heal. Aff. Forefr.*, 2024.

[7] E. Hassan and C. E. Omenogor, "AI powered predictive healthcare: Deep learning for early diagnosis, personalized treatment, and disease prevention," *Int. J. Sci. Res. Arch.*, vol. 14, no. 3, pp. 806–823, 2025, doi: 10.30574/ijsra.2025.14.3.0731.

[8] M. A. Islam *et al.*, "Harnessing Predictive Analytics: The Role of Machine Learning in Early Disease Detection and Healthcare Optimization," *J. Ecohumanism*, vol. 4, no. 3, pp. 312–321, 2025, doi: 10.62754/joe.v4i3.6642.

[9] S. Rani *et al.*, "Machine Learning-Powered Smart Healthcare Systems in the Era of Big Data: Applications, Diagnostic Insights, Challenges, and Ethical Implications," *Diagnostics*, vol. 15, no. 15, p. 1914, 2025, doi: 10.3390/diagnostics15151914.

[10] A. Anderies, J. A. R. W. Tchin, P. H. Putro, Y. P. Darmawan, and A. A. S. Gunawan, "Prediction of Heart Disease UCI Dataset

Using Machine Learning Algorithms," *Eng. Math. Comput. Sci. J.*, vol. 4, no. 3, pp. 87–93, 2022, doi: 10.21512/emacsjournal.v4i3.8683.

[11] Y. Yang, Y. Wang, Y. Li, S. Sen, L. Li, and Q. Liu, "Unleashing the Potential of Large Language Models," 2025, doi: https://doi.org/10.48550/arXiv.2402.17564.

[12] N. Absar *et al.*, "The Efficacy of Machine-Learning-Supported Smart System for Heart Disease Prediction," *Healthc.*, vol. 10, no. 6, pp. 1–19, 2022, doi: 10.3390/healthcare10061137.

[13] M. Phongying and S. Hiriote, "Diabetes Classification Using Machine Learning Techniques," *Computation*, vol. 11, no. 5, 2023, doi: 10.3390/computation11050096.

[14] M. El-Melegy *et al.*, "Prostate Cancer Diagnosis via Visual Representation of Tabular Data and Deep Transfer Learning," *Bioengineering*, vol. 11, no. 7, pp. 1–25, 2024, doi: 10.3390/bioengineering11070635.

[15] S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang, and D. A. Sontag, "TabLLM: Few-shot Classification of Tabular Data with Large Language Models. BT - International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain.," vol. 206, pp. 5549–5581, 2023, [Online]. Available: https://proceedings.mlr.press/v206/hegselmann23a.html

[16] W. Ren, T. Zhao, Y. Huang, and V. Honavar, "Deep Learning within Tabular Data: Foundations, Challenges, Advances and Future Directions," 2025, [Online]. Available: http://arxiv.org/abs/2501.03540

[17] T. Wahyuningsih, D. Manongga, I. Sembiring, and S. Wijono, "Comparison of Effectiveness of Logistic Regression, Naive Bayes, and Random Forest Algorithms in Predicting Student Arguments," *Procedia Comput. Sci.*, vol. 234, pp. 349–356, 2024, doi: 10.1016/j.procs.2024.03.014.

[18] A. I. Putri *et al.*, "Implementation of K-Nearest Neighbors, Naïve Bayes Classifier, Support Vector Machine and Decision Tree Algorithms for Obesity Risk Prediction," *Public Res. J. Eng. Data Technol. Comput. Sci.*, vol. 2, no. 1, pp. 26–33, 2024, doi: 10.57152/predatecs.v2i1.1110.

[19] J. Pasaribu, N. Yudistira, and W. Firdaus Mahmudy, "Tabular Data Classification and Regression: XGBoost or Deep Learning

with Retrieval-Augmented Generation," *IEEE Access*, vol. 12, no. November, pp. 191719–191732, 2024, doi: 10.1109/ACCESS.2024.3518205.

[20] M. Parmar and A. Tiwari, "Enhancing Text Classification Performance using Stacking Ensemble Method with TF-IDF Feature Extraction," *Int. Conf. Mob. Comput. Sustain. Informatics (ICMCSI), Lalitpur, Nepal*, pp. 166–174, 2024, doi: https://doi.org/10.1109/ICMCSI61536.2024.00031.

[21] P. M. T. Cortesão, "Towards Generalisation In Tabular Models With LLM-Learned Concepts," no. July, 2024.

[22] M. Jayawardhana *et al.*, "Transformers Boost the Performance of Decision Trees on Tabular Data across Sample Sizes," 2025, [Online]. Available: http://arxiv.org/abs/2502.02672

[23] M. Salsabil, N. Lutvi, and A. Eviyanti, "Implementasi Data Mining Dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest Dan Xgboost," *J. Ilm. Komputasi*, vol. 23, no. 1, pp. 51–58, 2024, doi: 10.32409/jikstik.23.1.3507.

[24] K. Ono and S. A. Lee, "Text Serialization and Their Relationship with the Conventional Paradigms of Tabular Machine Learning," 2024, [Online]. Available: http://arxiv.org/abs/2406.13846

[25] M. I. T. Csail, "TabLLM: Few-shot Classification of Tabular Data with Large Language Models," vol. 206, 2023, doi: https://doi.org/10.48550/arXiv.2210.10723.