Received: 13 Jun 2025 | Revised: 15 Apr 2026 | Accepted: 30 Apr 2025 | Available online: 13 May 2026
https://doi.org/10.13057/ijas.v9i1.103994

E-ISSN: 2621-086X
https://jurnal.uns.ac.id/ijas

Indonesian Journal of Applied Statistics Vol. 9, No.1, pp. 30-42, 2026

Bayesian Neural Network untuk Prediksi Diabetes: *Uncertainty Quantification* dalam *Machine Learning*

Sabrina Adnin Kamila, Kusman Sadik, Cici Suhaeni, Agus Mohamad Soleh
Program Studi Statistika dan Sains Data, IPB University, Bogor, Indonesia
*Corresponding Author. E-mail: sabrinaadnin@apps.ipb.ac.id

Abstrak

Penelitian ini bertujuan mengevaluasi dan membandingkan kinerja tiga model machine learning, yaitu random forest (RF), feedforward neural network (FNN), dan bayesian neural network (BNN), dalam klasifikasi diabetes menggunakan Diabetes Health Indicators Dataset dari UCI Machine Learning Repository yang memiliki ketidakseimbangan kelas. Prapemrosesan data meliputi normalisasi fitur menggunakan StandardScaler dan penanganan ketidakseimbangan kelas dengan synthetic minority over-sampling technique (SMOTE). Evaluasi model dilakukan menggunakan metrik akurasi dan skor F1, yang didukung oleh classification report dan confusion matrix. Hasil evaluasi menunjukkan bahwa RF menghasilkan akurasi tinggi (0,8493) namun skor F1 yang rendah (0,3386), yang mengindikasikan rendahnya sensitivitas model terhadap kasus positif diabetes. FNN memberikan performa yang lebih seimbang dengan skor F1 sebesar 0,4490 setelah penyesuaian threshold optimal. Sementara itu, BNN mencapai akurasi 0,8498 dan skor F1 sebesar 0,4043, serta memiliki keunggulan tambahan berupa kemampuan mengukur ketidakpastian prediksi melalui pendekatan Monte Carlo Dropout. Dengan demikian, FNN lebih unggul dalam keseimbangan klasifikasi, sementara BNN lebih relevan untuk aplikasi medis yang membutuhkan informasi tingkat kepercayaan prediksi guna mendukung pengambilan keputusan klinis yang lebih andal.

This study aims to evaluate and compare the performance of three machine learning models, namely random forest (RF), feedforward neural network (FNN), and bayesian neural network (BNN), for diabetes classification using the Diabetes Health Indicators Dataset from the UCI Machine Learning Repository, which exhibits significant class imbalance. Data preprocessing includes feature normalization using StandardScaler and class imbalance handling through synthetic minority over-sampling technique (SMOTE). Model performance is evaluated using accuracy and F1-score metrics, supported by classification report and confusion matrix analysis. The results show that RF achieves high accuracy (0.8493) but a low F1-score (0.3386), indicating poor sensitivity to positive diabetes cases. FNN provides more balanced performance with an F1-score of 0.4490 after optimal threshold adjustment. Meanwhile, BNN achieves an accuracy of 0.8498 and F1-score of 0.4043, while offering the additional advantage of uncertainty quantification through Monte Carlo Dropout. Therefore, FNN is more effective for balanced classification performance, while BNN is more suitable for medical applications that require prediction confidence information to support more reliable and informed clinical decision-making.

Kata Kunci: Prediksi diabetes, kuantifikasi ketidakpastian, bayesian neural network, classification imbalance, machine learning.

Keywords: *Diabetes prediction, uncertainty quantification, bayesian neural network, classification imbalance, machine learning.*

How to Cite:
S. A. Kamila, K. Sadik, C. Suhaeni, and A. M. Soleh, "Bayesian neural network untuk prediksi diabetes: uncertainty quantification dalam machine learning," *Indonesian Journal of Applied Statistics,* vol. 9, no. 1, pp. 1-15, 2026, doi: 10.13057/ijas.v9i1.103994.

1. PENDAHULUAN

Perkembangan kecerdasan buatan (*artificial intelligence*/AI) telah membawa transformasi signifikan dalam berbagai sektor, termasuk bidang kesehatan. *Machine learnig* (ML) sebagai salah satu cabang utama AI telah banyak dimanfaatkan untuk menganalisis data medis dan mendukung pengambilan keputusan klinis. Model ML mampu mengidentifikasi pola kompleks dalam data kesehatan yang sulit dikenali oleh pendekatan konvensional, seperti dalam deteksi dini penyakit dan prediksi hasil pemeriksaan medis [1]. Selain itu, penerapan AI melalui simulasi berbasis ML dan analisis data dalam pendidikan medis telah terbukti efektif dalam meningkatkan kemampuan klinis mahasiswa kedokteran [2].

Sebagian besar model ML yang digunakan saat ini masih bersifat deterministik, yaitu menghasilkan satu nilai prediksi tanpa mempertimbangkan ketidakpastian yang melekat pada data maupun parameter model [3]. Pendekatan deterministik ini berpotensi menimbulkan *overconfidence* terhadap hasil prediksi, terutama ketika dihadapkan pada data yang tidak terdistribusi secara serupa dengan data pelatihan (*out-of-distribution*) [4]. Dalam konteks medis, kondisi tersebut menjadi permasalahan krusial karena kesalahan prediksi dapat berdampak langsung terhadap keselamatan pasien.

Uncertainty quantification (UQ) merupakan komponen penting dalam pengembangan model ML untuk mengatasi keterbatasan pendekatan deterministik. Penerapan UQ memungkinkan pengukuran tingkat kepercayaan terhadap hasil prediksi, sehingga model tidak hanya memberikan *output* deterministik, tetapi juga informasi probabilistik yang dapat mendukung pengambilan keputusan klinis yang lebih andal dan terukur [5]. UQ tidak hanya mengukur tingkat ketidakpastian, tetapi juga mengidentifikasi sumbernya, baik yang berasal dari data (*aleatoric uncertainty*) maupun dari model (*epistemic uncertainty*). Dalam konteks medis, kemampuan ini penting untuk mengurangi kesalahan diagnosis akibat prediksi yang terlalu percaya diri [6]. Sebagai contoh, model deterministik dalam prediksi kondisi pasien pasca operasi cenderung mengabaikan variabilitas karakteristik pasien, seperti komorbiditas atau kondisi langka yang dapat memengaruhi akurasi prediksi.

Salah satu pendekatan yang dapat digunakan untuk mengintegrasikan UQ ke dalam model ML adalah *bayesian neural network* (BNN). Berbeda dengan jaringan saraf konvensional yang bersifat deterministik, BNN menggunakan pendekatan probabilistik dengan memodelkan parameter sebagai distribusi peluang. Hal ini memungkinkan model untuk menghasilkan distribusi prediksi serta mengukur tingkat ketidakpastian secara eksplisit. Dalam konteks medis, kemampuan ini menjadi sangat penting karena model tidak hanya memberikan hasil klasifikasi, tetapi juga informasi mengenai tingkat keyakinan terhadap prediksi tersebut sehingga dapat membantu mengidentifikasi kasus dengan risiko tinggi [7].

Penelitian ini menggunakan *Diabetes Health Indicators Dataset* yang berasal dari *UCI Machine Learning Repository*. Data ini terdiri dari 21 variabel yang mencakup indikator kesehatan dan gaya hidup yang berhubungan dengan risiko diabetes, seperti indeks massa tubuh (*body mass index*), tekanan darah tinggi, kolesterol tinggi, riwayat merokok, aktivitas fisik, serta konsumsi buah dan sayuran. Keberagaman variabel tersebut memungkinkan analisis yang lebih komprehensif terhadap faktor risiko diabetes sehingga data ini relevan untuk pengembangan model berbasis BNN yang tidak hanya mampu melakukan prediksi, tetapi juga mengkuantifikasi ketidakpastian hasil prediksi secara probabilistik.

2. METODE

2.1. Data

Penelitian ini menggunakan dataset *Diabetes Health Indicators* dari *UCI Machine Learning Repository*. Data ini mengandung berbagai indikator kesehatan untuk mengklasifikasikan individu ke dalam kategori diabetes atau sehat (1 = memiliki diabetes, 0 = tidak memiliki diabetes). Data ini memiliki Jumlah observasi sejumlah 253.680 sampel dan 21 variabel. Deskripsi variabel penelitian disajikan pada Tabel 1.

https://doi.org/10.13057/ijas.v9i1.103994
E-ISSN: 2621-086X

Tabel 1. Variabel penelitian

Variabel	Deskripsi
HighBP	Tekanan darah tinggi (1 = Ya, 0 = Tidak)
HighChol	Kolesterol tinggi (1 = Ya, 0 = Tidak)
CholCheck	Pemeriksaan kadar kolesterol (1 = Ya, 0 = Tidak)
BMI	Indeks massa tubuh
Smoker	Riwayat merokok (1 = Ya, 0 = Tidak)
Stroke	Riwayat stroke (1 = Ya, 0 = Tidak)
HeartDiseaseorAttack	Riwayat penyakit jantung (1 = Ya, 0 = Tidak)
PhysActivity	Aktivitas fisik rutin (1 = Ya, 0 = Tidak)
Fruits	Konsumsi buah setiap hari (1 = Ya, 0 = Tidak)
Veggies	Konsumsi sayuran setiap hari (1 = Ya, 0 = Tidak)
HvyAlcoholConsump	Konsumsi alkohol berlebihan (1 = Ya, 0 = Tidak)
AnyHealthCare	Akses perlindungan kesehatan (1 = Ya, 0 = Tidak)
NoDocbcCost	Hambatan ekonomi layanan kesehatan (1 = Ya, 0 = Tidak)
GenHlth	Kondisi kesehatan (1 = Sempurna, 2 = Sangat baik, 3 = Baik, 4 = Cukup, 5 = Buruk)
MentHlth	Jumlah hari mengalami gangguan kesehatan mental
PhysHlth	Jumlah hari mengalami gangguan kesehatan fisik
DiffWalk	Kesulitan berjalan atau menaiki tangga (1 = Ya, 0 = Tidak)
Sex	Jenis kelamin (1 = Laki-laki, 0 = Perempuan)
Age	Tingkat kelompok usia (1 = 18-24 tahun, 9 = 60-64 tahun, 13 = 80 tahun atau lebih)
Education	Tingkat pendidikan terakhir (1 = Tidak pernah sekolah atau hanya taman kanak-kanak, 2 = Sekolah Dasar, 3 = Sekolah Menengah, 4 = Lulus Sekolah Menengah, 5 = Kuliah, 6 = Lulusan Perguruan Tinggi)
Income	Tingkat pendapatan (1 = < $10.000, 5 = < $35.000, 8 = ≥ $75.000)

2.2. *Data Preprocessing*

Data preprocessing merupakan langkah penting dalam *machine learning* yang bertujuan untuk meningkatkan kualitas data sebelum digunakan dalam pemodelan [8]. Langkah-langkah *data preprocessing* meliputi pembersihan data, transformasi fitur, penyeimbangan kelas, dan pembagian data.

Tahap pertama dalam *data preprocessing* adalah pembersihan data yang mencakup deteksi dan penanganan nilai yang hilang (*missing values*). Nilai yang hilang dapat diimputasi menggunakan metode seperti *mean/median imputation*. Selanjutnya, dilakukan transformasi fitur agar data dapat diproses dengan baik oleh model *machine learning*. Transformasi ini meliputi normalisasi atau standardisasi fitur numerik untuk menyamakan skala variabel yang berbeda [9].

Jika data memiliki ketidakseimbangan kelas (*class imbalance*), maka akan menggunakan teknik *oversampling, undersampling,* atau *synthetic minority oversampling technique* (SMOTE) untuk meningkatkan jumlah sampel di kelas minoritas [10]. Data kemudian dibagi menjadi 80% digunakan sebagai data latih dan 20% digunakan sebagai data uji yang bertujuan untuk memastikan bahwa proporsi kelas pada data pengujian tetap seimbang.

2.3. *Random Forest* (RF)

Random forest (RF) merupakan metode klasifikasi berbasis *ensemble learning* yang menggabungkan sejumlah pohon keputusan (*decision tree*) yang dibangun secara independen. Prediksi akhir diperoleh

https://doi.org/10.13057/ijas.v9i1.103994
E-ISSN: 2621-086X

melalui mekanisme *majority voting* dari seluruh pohon yang terbentuk sehingga mampu meningkatkan akurasi dan stabilitas model [11]. RF merupakan pengembangan dari teknik *bagging* (*bootstrap aggregating*). Pada pendekatan ini, setiap pohon dilatih menggunakan subset data yang diperoleh melalui *bootstrap sampling*. Selain itu, dalam proses pembentukan pohon, RF menerapkan pemilihan subset fitur secara acak pada setiap titik pemisahan (*split*). Strategi tersebut bertujuan untuk mengurangi korelasi antar pohon sehingga dapat menurunkan variansi model dan meningkatkan generalisasi [11],[12]. RF memiliki keunggulan dalam mengatasi *overfitting* serta mampu menangani data dengan dimensi tinggi dibandingkan dengan metode pohon keputusan tunggal. Hal ini menjadikan RF sebagai salah satu metode *baseline* yang kuat dalam berbagai permasalahan klasifikasi, termasuk pada data medis yang kompleks [13].

Dalam penelitian ini, beberapa *hyperparameter* utama yang digunakan dalam model *random forest* meliputi [11]:

1. Jumlah pohon (n_estimators): semakin banyak pohon maka semakin stabil prediksi, namun biaya komputasi meningkat.
2. Kedalaman maksimum pohon (max_depth): parameter ini membatasi kompleksitas tiap pohon untuk mencegah *overfitting*.
3. Jumlah fitur yang dipertimbangkan pada setiap *split* (max_features): berpengaruh terhadap variasi antar pohon.
4. Kriteria pemisahan (criterion): *gini* atau *entropy* untuk klasifikasi dan MSE atau MAE untuk regresi.

2.4. *Feedforward Neural Network* (FNN)

Feedforward neural network (FNN) merupakan salah satu bentuk dasar jaringan saraf tiruan yang terdiri dari lapisan *input,* satu atau lebih lapisan tersembunyi (*hidden layer*), dan satu lapisan *output*. Pada FNN, aliran informasi bergerak secara searah dari *input* menuju *output* tanpa adanya umpan balik (*feedback*) atau siklus [14]. Setiap neuron melakukan operasi matematis berupa perkalian antara *input* dan bobot, penambahan bias, serta penerapan fungsi aktivasi nonlinier, seperti *sigmoid, tanh,* atau *rectified linear unit* (ReLU) untuk menghasilkan *output* [15].

Proses pembelajaran pada FNN dilakukan melalui algoritma *backpropagation* yang melibatkan tiga tahap utama, yaitu *forward propagation* untuk menghitung *output,* perhitungan *error* terhadap nilai target, dan *backward propagation* untuk memperbarui bobot berdasarkan gradien *error*. Proses pelatihan ini umumnya menggunakan fungsi *loss* seperti *mean squared error* (MSE) atau *binary cross-entrophy,* serta algoritma optimasi seperti *stochastic gradient descent* (SGD) atau *adam optimizer* untuk mempercepat konvergensi model [15],[16].

FNN memiliki keterbatasan fundamental karena bersifat deterministik. Bobot jaringan yang dihasilkan selama pelatihan bersifat tetap, sehingga model hanya menghasilkan satu nilai prediksi tanpa informasi mengenai ketidakpastian (*uncertainty*). Hal ini berpotensi menimbulkan prediksi yang *overconfident* khususnya pada data yang berada di luar distribusi pelatihan (*out-of-distribution*) [14].

2.5. *Bayesian Neural Network* (BNN)

Bayesian neural network (BNN) merupakan pengembangan dari jaringan saraf tiruan yang mengintegrasikan pendekatan *probabilistic inference* sehingga model tidak hanya menghasilkan prediksi, tetapi juga mampu memberikan estimasi ketidakpastian dari hasil prediksi tersebut. Pendekatan ini menjadi penting dalam domain medis karena dapat memberikan informasi tambahan terkait tingkat kepercayaan terhadap keputusan yang dihasilkan model.

Secara teori, model Bayesian bekerja berdasarkan teorema bayes yang dinyatakan sebagai berikut [14]:

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)} \quad (1)$$

https://doi.org/10.13057/ijas.v9i1.103994
E-ISSN: 2621-086X

di mana:

$p(w|D)$: *posterior distribution* dari parameter bobot w setelah melihat data D

$p(D|w)$: *likelihood* dari data yang diamati

$p(w)$: *prior* atas parameter sebelum pengamatan data

$p(D)$: *evidence* atau probabilitas dari data

Berbeda dengan jaringan saraf konvensional yang mengoptimasi bobot sebagai nilai deterministik melalui *backpropagation,* BNN memodelkan bobot sebagai variabel acak yang mengikuti distribusi probabilistik. Pendekatan ini memungkinkan model untuk menangkap *epistemic uncertainty,* yaitu ketidakpastian yang disebabkan oleh keterbatasan data atau struktur model yang belum optimal [14].

Dalam inferensi bayesian, digunakan aproksimasi seperti *variational inference* (VI) yang bertujuan mencari distribusi aproksimasi $q(w)$ yang mendekati distribusi *posterior* sebenarnya $p(w|D)$. Optimasi dilakukan dengan meminimalkan nilai Kullback-Leibler (KL) *divergence,* yang mengukur jarak antara dua distribusi. Pendekatan VI digunakan karea perhitungan *posterior* bersifat eksak dan jarang digunakan dalam analisis jaringan saraf berukuran besar [14],[17].

Selain itu, implementasi BNN dalam penelitian ini menggunakan pendekatan *monte carlo dropout* (MC *dropout*) untuk mengestimasi ketidakpastian prediksi. Pada metode ini, mekanisme *dropout* tetap diaktifkan selama tahap inferensi, sehingga model menghasilkan beberapa sampel prediksi melalui proses *forward pass* berulang. Rata-rata dari prediksi tersebut digunakan sebagai *output* akhir, sedangkan varians atau deviasi standar digunakan sebagai ukuran ketidakpastian [18].

2.6. Evaluasi Model

Evaluasi model dilakukan untuk mengukur kinerja model dalam melakukan klasifikasi data diabetes. Pada penelitian ini, evaluasi dilakukan menggunakan beberapa metrik, yaitu *accuracy* dan *F1-score*, yang dihitung berdasarkan hasil prediksi pada data uji [19].

Accuracy digunakan untuk mengukur proporsi jumlah prediksi yang benar terhadap seluruh data, yang dirumuskan sebagai berikut [20]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

di mana *true positive* (TP) merupakan jumlah data positif yang diprediksi benar, *true negative* (TN) merupakan jumlah data negatif yang diprediksi benar, *false positive* (FP) merupakan jumlah data negatif yang diprediksi sebagai positif, dan *false negative* (FN) merupakan jumlah data positif yang diprediksi sebagai negatif.

Selain itu, digunakan *F1-score* untuk mengevaluasi keseimbangan antara *precision* dan *recall*, khususnya pada data dengan distribusi kelas yang tidak seimbang. *F1-score* dirumuskan sebagai berikut [21]:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

dengan *precision* dan *recall* masing-masing didefinisikan sebagai:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

Penggunaan *F1-score* dalam penelitian ini bertujuan untuk memberikan evaluasi yang lebih representatif terhadap performa model dalam mendeteksi kelas minoritas, yaitu pasien dengan diabetes. Hal ini penting karena ketidakseimbangan data dapat menyebabkan metrik *accuracy* menjadi kurang informatif [19].

https://doi.org/10.13057/ijas.v9i1.103994
E-ISSN: 2621-086X

Selain metrik kuantitatif, evaluasi model juga dilakukan menggunakan *confusion matrix* untuk memberikan gambaran distribusi prediksi model terhadap masing-masing kelas [22]. Melalui *confusion matrix*, dapat dianalisis kesalahan klasifikasi yang terjadi, khususnya pada kasus *false negative* yang memiliki implikasi penting dalam konteks medis.

3. HASIL DAN PEMBAHASAN

3.1. Kualitas dan Kelayakan Data

Tahap awal analisis dilakukan dengan mengevaluasi kualitas data melalui pemeriksaan nilai yang hilang (*missing values*). Proses ini dilakukan menggunakan fungsi *df.isnull().sum()* pada Python (*library Pandas*) yang berfungsi untuk menghitung jumlah nilai kosong pada setiap variabel dalam data. Berdasarkan hasil pemeriksaan tersebut, seluruh variabel dalam data tidak mengandung nilai yang hilang. Hal ini menunjukkan bahwa data telah memiliki kualitas yang baik dan dapat langsung digunakan dalam proses pemodelan tanpa memerlukan teknik imputasi. Ketiadaan *missing value* juga mengurangi potensi bias yang dapat muncul akibat asumsi dalam proses imputasi sehingga performa model yang dihasilkan lebih merepresentasikan pola asli dalam data.

3.2. Normalisasi dan Pembagian Data

Setelah dilakukan pemeriksaan kualitas data dan dipastikan tidak terdapat nilai yang hilang, tahap selanjutnya adalah normalisasi fitur dan pembagian dataset. Seluruh fitur numerik dinormalisasi menggunakan metode *StandardScaler*, yang mentransformasikan data sehingga memiliki rata-rata sebesar 0 dan standar deviasi sebesar 1. Proses ini bertujuan untuk menyamakan skala antar variabel sehingga tidak ada fitur yang mendominasi proses pembelajaran model akibat perbedaan rentang nilai.

Normalisasi merupakan langkah penting terutama pada model berbasis jaringan saraf seperti FNN dan BNN, yang sensitif terhadap skala input. Dengan data yang telah dinormalisasi, proses optimasi menjadi lebih stabil dan konvergensi model dapat dicapai dengan lebih efisien. Selanjutnya, dataset dibagi menjadi dua subset, yaitu data latih (*training set*) sebesar 80% dan data uji (*testing set*) sebesar 20%. Pembagian dilakukan menggunakan teknik *stratified splitting* untuk memastikan bahwa proporsi kelas pada variabel target tetap terjaga di kedua subset. Hal ini penting untuk menghindari bias dalam evaluasi model, terutama pada kondisi data yang tidak seimbang.

3.3. Distribusi Kelas dan Penanganan Ketidakseimbangan

Setelah dilakukan proses normalisasi dan pembagian data, analisis terhadap distribusi variabel target pada data latih menunjukkan adanya ketidakseimbangan kelas (*class imbalance*) yang cukup signifikan. Variabel target dalam penelitian ini dikodekan sebagai 0 untuk individu yang tidak mengidap diabetes dan 1 untuk individu yang mengidap diabetes. Berdasarkan hasil eksplorasi, jumlah observasi pada kelas negatif (0) jauh lebih besar dibandingkan kelas positif (1), yaitu sebesar 174.667 observasi pada kelas 0 dan 28.277 observasi pada kelas 1.

Ketidakseimbangan ini berpotensi menyebabkan model pembelajaran mesin cenderung bias terhadap kelas mayoritas, sehingga model lebih sering memprediksi kelas negatif dan mengabaikan pola penting pada kelas positif. Dalam konteks medis, kondisi ini dapat meningkatkan risiko terjadinya *false negative* (FN), yaitu kasus diabetes yang tidak terdeteksi oleh model.

Untuk mengatasi permasalahan tersebut, diterapkan metode *synthetic minority over-sampling technique* (SMOTE) pada data latih. Metode ini bekerja dengan membangkitkan sampel sintetis pada kelas minoritas berdasarkan kedekatan antar data dalam ruang fitur, sehingga tidak hanya meningkatkan jumlah data, tetapi juga memperkaya variasi pola pada kelas minoritas. Setelah penerapan SMOTE, distribusi kelas menjadi seimbang dengan jumlah observasi sebesar 174.667 pada masing-masing kelas.

Meskipun SMOTE efektif dalam meningkatkan representasi kelas minoritas, metode ini juga memiliki keterbatasan. Pembentukan data sintetis yang terlalu menyerupai data asli berpotensi

https://doi.org/10.13057/ijas.v9i1.103994
E-ISSN: 2621-086X

menyebabkan model mengalami overfitting, terutama jika variasi data yang dihasilkan tidak sepenuhnya mencerminkan distribusi sebenarnya. Oleh karena itu, evaluasi model pada data uji tetap diperlukan untuk memastikan kemampuan generalisasi model.

3.4. Performa Model *Random Forest*

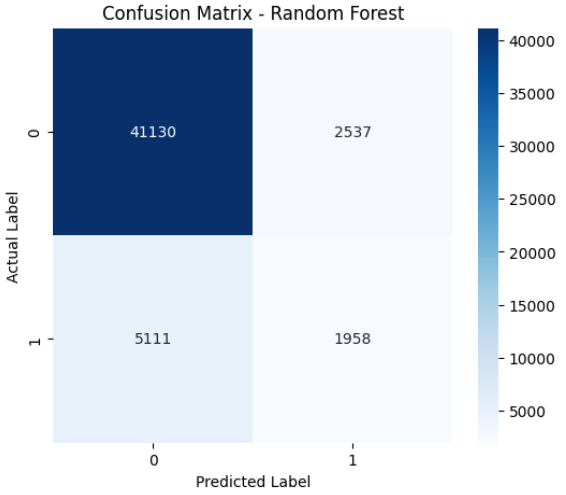
Model *random forest* pada penelitian ini dibangun menggunakan parameter *default* dari pustaka *scikit-learn*, dengan jumlah pohon (*n_estimators*) sebanyak 100, sementara parameter lain seperti *max_depth* dan *max_features* menggunakan nilai bawaan. Pendekatan ini digunakan untuk mengevaluasi performa baseline model tanpa proses optimasi hyperparameter lebih lanjut.

Berdasarkan hasil evaluasi pada data uji, model *random forest* menghasilkan akurasi sebesar 0,8493 dan *F1-score* sebesar 0,3386. Nilai akurasi yang tinggi menunjukkan bahwa model mampu mengklasifikasikan sebagian besar observasi dengan benar. Namun demikian, *F1-score* yang relatif rendah mengindikasikan bahwa performa model tidak seimbang dalam mendeteksi kedua kelas. Untuk memperoleh gambaran yang lebih rinci mengenai performa model, disajikan *classification report* pada Tabel 2.

Table 2. *Classification report* model *random forest*

Kelas	*Precision*	*Recall*	*F1-score*	*Support*
0 (Tidak Diabetes)	0,89	0,94	0,91	43667
1 (Diabetes)	0,44	0,28	0,34	7069
Accuracy			0,85	50736
Macro Avg	0,66	0,61	0,63	50736

Berdasarkan Tabel 2, terlihat bahwa nilai *recall* untuk kelas positif (diabetes) hanya sebesar 0,28, yang menunjukkan bahwa sebagian besar kasus diabetes tidak berhasil terdeteksi oleh model. Hal ini berimplikasi pada tingginya jumlah *false negative* (FN), yang dalam konteks medis merupakan kesalahan yang krusial karena dapat menyebabkan pasien tidak mendapatkan penanganan yang tepat. Untuk memperjelas distribusi kesalahan klasifikasi, confusion matrix model Random Forest disajikan pada Gambar 1.



Gambar 1. *Confusion matrix* model *random forest*

Berdasarkan *confusion matrix* tersebut, terlihat bahwa jumlah prediksi pada kelas negatif jauh lebih dominan dibandingkan kelas positif. Hal ini menunjukkan bahwa model cenderung mengklasifikasikan data ke dalam kelas mayoritas, meskipun data telah diseimbangkan pada tahap pelatihan. Fenomena ini dapat dijelaskan melalui mekanisme *majority voting* pada *random forest*. Dalam kondisi data yang tidak seimbang, setiap pohon keputusan dalam ensemble cenderung lebih sering mempelajari pola dari kelas

https://doi.org/10.13057/ijas.v9i1.103994
E-ISSN: 2621-086X

mayoritas, sehingga keputusan akhir model menjadi bias terhadap kelas tersebut. Akibatnya, model mengalami kesulitan dalam membentuk batas keputusan (*decision boundary*) yang optimal untuk mengidentifikasi kelas minoritas. Selain itu, meskipun SMOTE telah digunakan untuk menyeimbangkan data latih, pendekatan ini tidak selalu menjamin bahwa model mampu memahami karakteristik kompleks dari kelas minoritas. Hal ini terutama terjadi jika data sintetis yang dihasilkan tidak sepenuhnya merepresentasikan distribusi sebenarnya.

Perbedaan yang cukup signifikan antara nilai akurasi dan *F1-score* menegaskan bahwa penggunaan akurasi sebagai satu-satunya metrik evaluasi tidak cukup representatif pada data yang tidak seimbang. Oleh karena itu, metrik seperti *F1-score* dan *recall* menjadi lebih relevan dalam mengevaluasi kemampuan model dalam mendeteksi kasus positif. Secara keseluruhan, hasil ini menunjukkan bahwa meskipun *random forest* memiliki performa yang baik dalam hal akurasi, model ini memiliki keterbatasan dalam mendeteksi kelas minoritas. Oleh karena itu, diperlukan pendekatan model lain yang lebih mampu meningkatkan sensitivitas terhadap kelas positif, yang akan dibahas pada bagian selanjutnya.

Hasil yang diperoleh pada model *random forest* ini sejalan dengan penelitian Thomas dan Kaliraj [13] yang melaporkan bahwa *random forest* mampu memberikan performa klasifikasi yang lebih baik dibandingkan beberapa metode lain pada data terstruktur, terutama karena kemampuannya dalam menangani data berdimensi tinggi dan mengurangi *overfitting*. Namun, pada data dengan ketidakseimbangan kelas, performa model tetap perlu dievaluasi secara hati-hati meskipun teknik SMOTE telah diterapkan. Kondisi ini sejalan dengan penelitian Kocak, dkk. [20] yang menyatakan bahwa pada data tidak seimbang, akurasi dapat memberikan gambaran performa yang terlalu optimistis karena model cenderung mengikuti kelas mayoritas. Oleh karena itu, metrik seperti *F1-score* lebih representatif dalam mengevaluasi kemampuan model dalam mengidentifikasi kelas minoritas.

3.5. Performa Model *Feedforward Neural Network* (FNN)

Model *feedforward neural network* (FNN) digunakan sebagai pendekatan lanjutan untuk mengatasi keterbatasan model *random forest* dalam menangani pola *non-linear* dan meningkatkan sensitivitas terhadap kelas minoritas. Model FNN dibangun menggunakan arsitektur jaringan saraf sederhana yang terdiri dari dua lapisan tersembunyi (*hidden layer*) dengan fungsi aktivasi *rectified linear unit* (ReLU), serta satu lapisan *output* dengan fungsi aktivasi sigmoid untuk menghasilkan probabilitas klasifikasi biner.

Berdasarkan hasil evaluasi awal menggunakan *threshold default* sebesar 0,5, model FNN menghasilkan akurasi sebesar 0,7730 dan *F1-score* sebesar 0,4467. Hasil ini menunjukkan bahwa model telah mampu meningkatkan keseimbangan performa dibandingkan *random forest*, terutama dalam mendeteksi kelas positif [14]. Namun demikian, penggunaan *threshold default* belum tentu memberikan kombinasi terbaik antara *precision* dan *recall*, khususnya pada data dengan distribusi kelas yang tidak seimbang.

Untuk memperoleh performa yang lebih optimal, dilakukan penyesuaian *threshold* klasifikasi. Hasil eksperimen menunjukkan bahwa *threshold* optimal diperoleh pada nilai 0,5425. Dengan menggunakan threshold ini, model FNN menghasilkan akurasi sebesar 0,7899 dan *F1-score* sebesar 0,4490. Meskipun peningkatan *F1-score* relatif tidak besar, hasil ini menunjukkan bahwa penyesuaian *threshold* mampu memberikan perbaikan performa yang lebih terarah dibandingkan penggunaan *threshold default*.

Table 3. Classification report model feedforward neural network

Kelas	*Precision*	*Recall*	*F1-score*	*Support*
0 (Tidak Diabetes)	0,93	0,82	0,87	43667
1 (Diabetes)	0,35	0,61	0,45	7069
Accuracy			0,79	50736
Macro Avg	0,64	0,72	0,66	50736

https://doi.org/10.13057/ijas.v9i1.103994
E-ISSN: 2621-086X

Berdasarkan Tabel 3, terlihat bahwa model FNN menunjukkan peningkatan kemampuan dalam mendeteksi kelas positif dibandingkan model *random forest*, yang tercermin dari nilai *recall* yang lebih tinggi. Hal ini menunjukkan bahwa model lebih sensitif dalam mengidentifikasi individu yang mengidap diabetes, meskipun terdapat *trade-off* dalam bentuk peningkatan jumlah *false positive*.

Penyesuaian *threshold* dari 0,5 menjadi 0,5425 mengindikasikan bahwa model menjadi lebih selektif dalam memberikan prediksi positif. Pendekatan ini membantu mengurangi prediksi positif yang tidak tepat, sekaligus mempertahankan kemampuan model dalam mendeteksi kasus diabetes secara cukup baik. Dengan demikian, *threshold* tuning menjadi langkah penting dalam meningkatkan keseimbangan performa model pada data yang tidak seimbang.

Secara keseluruhan, model FNN menunjukkan performa yang lebih seimbang dibandingkan *random forest*, khususnya dalam hal kemampuan mendeteksi kelas minoritas. Namun demikian, model ini masih memiliki keterbatasan karena bersifat deterministik, sehingga tidak mampu memberikan informasi mengenai tingkat ketidakpastian (*uncertainty*) dari prediksi yang dihasilkan [4].

3.6. Performa Model *Bayesian Neural Network* (BNN)

Sebagai pengembangan lebih lanjut dari model *feedforward neural network* (FNN), penelitian ini mengimplementasikan *bayesian neural network* (BNN) untuk mengatasi keterbatasan model deterministik dalam merepresentasikan ketidakpastian prediksi. Pada penelitian ini, pendekatan BNN diimplementasikan menggunakan teknik *monte carlo dropout*, di mana lapisan *dropout* tetap diaktifkan pada tahap inferensi untuk menghasilkan distribusi prediksi.

Arsitektur model BNN yang digunakan serupa dengan FNN, yaitu terdiri dari dua lapisan tersembunyi (*hidden layer*) dengan masing-masing 64 dan 32 neuron serta fungsi aktivasi ReLU. Perbedaan utama terletak pada penambahan lapisan *dropout* sebesar 0,1 pada setiap *hidden layer* yang tetap aktif saat proses prediksi. Model dilatih menggunakan optimizer Adam dengan fungsi *loss binary crossentropy* selama 20 *epoch*.

Untuk menghasilkan estimasi ketidakpastian, dilakukan sebanyak 100 kali proses prediksi (*monte carlo sampling*). Nilai prediksi akhir diperoleh dari rata-rata hasil prediksi, sedangkan standar deviasi digunakan sebagai ukuran ketidakpastian. Selanjutnya, dilakukan klasifikasi dengan menggunakan threshold sebesar 0,7 pada nilai probabilitas rata-rata. Pemilihan *threshold* yang lebih tinggi dari nilai *default* (0,5) bertujuan untuk meningkatkan tingkat kepercayaan model dalam memberikan prediksi positif.

Berdasarkan hasil evaluasi pada data uji, model BNN menghasilkan akurasi sebesar 0,8498 dan *F1-score* sebesar 0,4043. Nilai ini menunjukkan bahwa performa BNN berada di antara *random forest* dan FNN, dengan peningkatan kemampuan dalam mendeteksi kelas minoritas dibandingkan *random forest*, namun sedikit di bawah FNN dalam hal keseimbangan *F1-score*.

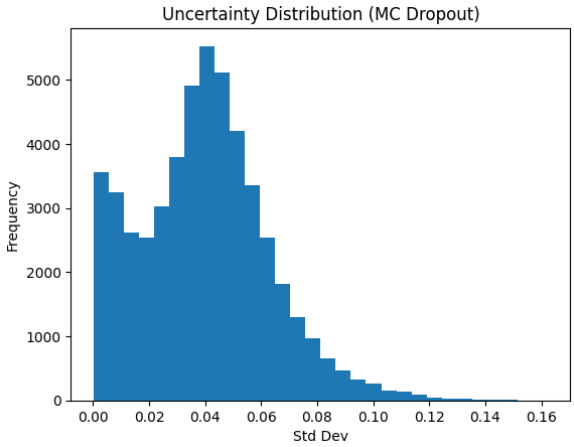
Table 4. *Classification report* model *bayesian neural network*

Kelas	*Precision*	*Recall*	*F1-score*	*Support*
0 (Tidak Diabetes)	0,90	0,93	0,91	43667
1 (Diabetes)	0,45	0,37	0,40	7069
Accuracy			0,85	50736
Macro Avg	0,68	0,65	0,66	50736

Berdasarkan Tabel 4, terlihat bahwa model BNN menunjukkan peningkatan kemampuan dalam mendeteksi kelas positif dibandingkan *random forest*, yang ditunjukkan oleh nilai *recall* sebesar 0,37 (dibandingkan 0,28 pada *random forest*). Hal ini menunjukkan bahwa pendekatan probabilistik pada BNN mampu meningkatkan sensitivitas model terhadap kasus diabetes. Namun demikian, peningkatan *recall* ini diikuti dengan penurunan precision yang relatif moderat, yang mengindikasikan adanya peningkatan jumlah *false positive*. Hal ini mencerminkan adanya *trade-off* antara kemampuan mendeteksi kasus positif dan akurasi prediksi.

https://doi.org/10.13057/ijas.v9i1.103994
E-ISSN: 2621-086X

Dibandingkan dengan FNN, performa BNN dalam hal *F1-score* masih sedikit lebih rendah. Hal ini dapat disebabkan oleh penggunaan *threshold* yang lebih tinggi (0,7), yang membuat model menjadi lebih konservatif dalam memberikan prediksi positif. Meskipun demikian, pendekatan ini membantu mengurangi prediksi positif yang tidak tepat (*false positive*).



Gambar 2. Distribusi ketidakpastian prediksi (*MC dropout*)

Distribusi ketidakpastian prediksi ditunjukkan melalui histogram standar deviasi hasil *monte carlo sampling*. Berdasarkan Gambar 2, terlihat bahwa sebagian besar prediksi memiliki nilai standar deviasi yang rendah (sekitar 0,02 hingga 0,06), yang menunjukkan bahwa model memiliki tingkat kepercayaan yang cukup tinggi terhadap sebagian besar prediksi. Meski demikian, terdapat sejumlah kecil prediksi dengan nilai ketidakpastian yang lebih tinggi yang dapat diinterpretasikan sebagai kasus ambigu atau sulit diklasifikasikan. Dalam konteks medis, informasi ini sangat penting karena memungkinkan identifikasi kasus yang memerlukan evaluasi lebih lanjut oleh tenaga medis.

Secara keseluruhan, model BNN tidak hanya memberikan performa klasifikasi yang kompetitif, tetapi juga menawarkan keunggulan tambahan dalam bentuk estimasi ketidakpastian prediksi. Hal ini menjadikan BNN lebih unggul dibandingkan model deterministik seperti *random forest* dan FNN dalam konteks aplikasi medis, di mana tingkat kepercayaan terhadap prediksi menjadi faktor yang krusial dalam pengambilan keputusan.

Hasil yang diperoleh pada model BNN ini konsisten dengan kajian Magris dan Losifidis [14] yang menegaskan bahwa pendekatan probabilistik pada *bayesian neural network* mampu memberikan informasi ketidakpastian prediksi secara lebih eksplisit dibandingkan model deterministik. Hasil kuantifikasi ketidakpastian pada penelitian ini juga menunjukkan bahwa sebagian besar prediksi model memiliki tingkat keyakinan yang relatif baik. Temuan ini sejalan dengan penelitian Whata, dkk. [18] yang menunjukkan bahwa pendekatan *monte carlo dropout* efektif digunakan untuk kuantifikasi ketidakpastian pada klasifikasi citra medis. Selain itu, penggunaan *threshold* yang lebih tinggi pada BNN dibandingkan FNN menunjukkan bahwa model Bayesian cenderung lebih selektif dalam menghasilkan prediksi positif. Secara keseluruhan, meskipun nilai *F1-score* BNN sedikit berada di bawah FNN, kemampuan BNN dalam menyediakan informasi probabilistik menjadikannya lebih relevan untuk skenario medis berisiko tinggi. Hal ini juga sejalan dengan pembahasan Rudner dan Toner [4] mengenai pentingnya *uncertainty quantification* untuk meningkatkan reliabilitas sistem *machine learning* pada aplikasi dunia nyata.

Table 5. Ringkasan evaluasi model

Model	Akurasi	*F1-Score*
Random Forest	0,8493	0,3386
Feedforward Neural Network	0,7899	0,4490
Bayesian Neural Network	0,8498	0,4043

https://doi.org/10.13057/ijas.v9i1.103994
E-ISSN: 2621-086X

Berdasarkan Tabel 5, terlihat secara ringkas bahwa BNN dan RF memiliki tingkat akurasi yang hampir setara dan lebih tinggi dibandingkan FNN, namun FNN unggul dalam *F1-score* sebagai metrik yang lebih relevan untuk data tidak seimbang. Hal ini menegaskan bahwa pemilihan model tidak dapat semata-mata didasarkan pada akurasi, melainkan harus mempertimbangkan konteks penggunaan dan metrik evaluasi yang tepat [19], [20].

4. SIMPULAN

Penelitian ini mengevaluasi kinerja model *random forest*, *feedforward neural network* (FNN), dan *bayesian neural network* (BNN) dalam klasifikasi diabetes pada data dengan ketidakseimbangan kelas. Hasil evaluasi menunjukkan bahwa *random forest* menghasilkan akurasi yang relatif tinggi (0,8493), namun memiliki *F1-score* yang rendah (0,3386), yang mengindikasikan keterbatasan model dalam mendeteksi kasus positif (diabetes). Model FNN memberikan performa yang lebih seimbang dengan *F1-score* sebesar 0,4490, meskipun akurasi yang dihasilkan lebih rendah dibandingkan *random forest*. Hal ini menunjukkan bahwa FNN lebih efektif dalam menangani data tidak seimbang, khususnya dalam meningkatkan kemampuan deteksi terhadap kelas minoritas.

Sementara itu, *bayesian neural network* menghasilkan akurasi sebesar 0,8498 dan *F1-score* sebesar 0,4043, yang menunjukkan performa yang kompetitif dibandingkan model lainnya. Meskipun *F1-score* BNN masih berada di bawah FNN, model ini memiliki keunggulan utama dalam menghasilkan estimasi ketidakpastian prediksi melalui pendekatan probabilistik berbasis *monte carlo dropout*. Temuan ini menunjukkan bahwa tidak terdapat satu model yang secara dominan unggul pada seluruh metrik evaluasi. FNN lebih optimal dalam hal keseimbangan performa klasifikasi, sedangkan BNN memberikan nilai tambah penting berupa informasi ketidakpastian prediksi yang tidak dimiliki oleh model deterministik.

Dalam konteks aplikasi medis, kemampuan untuk mengukur tingkat kepercayaan prediksi menjadi faktor krusial dalam mendukung pengambilan keputusan klinis. Oleh karena itu, penggunaan BNN menjadi lebih relevan pada skenario berisiko tinggi, di mana interpretasi terhadap ketidakpastian dapat membantu mengidentifikasi kasus yang memerlukan evaluasi lebih lanjut. Penelitian selanjutnya dapat mengembangkan pendekatan ini dengan mengeksplorasi metode bayesian yang lebih kompleks, seperti *variational inference* atau integrasi dengan teknik *calibration*, untuk meningkatkan keseimbangan antara performa klasifikasi dan kualitas estimasi ketidakpastian.

REFERENCES

[1] J. Bajwa, U. Munir, A. Nori, and B. Williams, "Artificial intelligence in healthcare: transforming the practice of medicine," *Future Healthcare Journal*, vol. 8, no. 2, pp. e188–e194, 2021, doi: https://doi.org/10.7861/fhj.2021-0095.

[2] S. A. Alowais, S. S. Alghamdi, N. Alsuhebany, T. Alqahtani, A. I. Alshaya, S. N. Almohareb, A. Aldairem, M. Alrashed, K. Bin Saleh, H. A. Badreldin, M. S. Al Yami, S. Al Harbi, and A. M. Albekairy, "Revolutionizing healthcare: the role of artificial intelligence in clinical practice," In *BMC Medical Education*, vol. 23, pp. 1-15, 2023, doi: https://doi.org/10.1186/s12909-023-04698-z.

[3] S. Liu, D. Lu, S. L. Painter, N. A. Griffiths, and E. M. Pierce, "Uncertainty quantification of machine learning models to improve streamflow prediction under changing climate and environmental conditions," *Frontiers in Water*, vol. 5, pp. 01-15, 2023, doi: https://doi.org/10.3389/frwa.2023.1150126.

[4] T. G. J. Rudner and H. Toner, *Issue Brief Key Concepts in AI Safety Reliable Uncertainty Quantification in Machine Learning*, 2024.

[5] K. A. Wahid, Z. Y. Kaffey, D. P. Farris, L. Humbert-Vidan, A. C. Moreno, M. Rasmussen, J. Ren, M. A. Naser, T. J. Netherton, S. Korreman, G. Balakrishnan, C. D. Fuller, D. Fuentes, and M. J. Dohopolski, "Artificial Intelligence Uncertainty Quantification in Radiotherapy Applications - A Scoping Review," *MedRxiv*, 2024, doi: https://doi.org/10.1101/2024.05.13.24307226.

[6] Y. Shi, P. Wei, K. Feng, D. -C. Feng, and M. Beer, "A survey on machine learning approaches for uncertainty quantification of engineering systems," *Machine Learning for Computational Science and Engineering*, vol. 1, no. 11, 2025, doi: https://doi.org/10.1007/s44379-024-00011-x.

[7] S. Ochella, F. Dinmohammadi, and M. Shafiee, "Bayesian neural networks for uncertainty quantification in remaining useful life prediction of systems with sensor monitoring," *Advances in Mechanical Engineering*, vol. 16, no. 7, 2024, doi: https://doi.org/10.1177/16878132241239802.

[8] P. Yasodha, "Data preprocessing methods for machine learning: An empirical comparison," *International Journal for Multidisciplinary Research*, vol. 7, no. 3, 2025, Available: https://www.ijfmr.com/papers/2025/3/48569.pdf.

[9] Y. D. Pratama and A. Salam, "Comparison of data normalization techniques on knn classification performance for pima indians diabetes dataset," *Journal of Applied Informatics and Computing (JAIC),* vol. 9, no. 3, 2025, Available: http://jurnal.polibatam.ac.id/index.php/JAIC.

[10] S. Fathmah, D. Kartini, F. Abadi, I. Budiman, and M. I. Mazdadi, "Implementation of PPCA imputation, SMOTE-N class balancing in hepatitis classification using naïve bayes," *Juita: Jurnal Informatika*, vol. 12, no. 2, pp. 169-176, 2024, doi: https://doi.org/10.30595/juita.v12i2.21528.

[11] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[12] O. Theobald, *Machine Learning for Absolute Beginners: A Plain English Introduction*. London, U.K.: Scatterplot Press, 2017.

[13] N. S. Thomas and S. Kaliraj, "An improved and optimized random forest based approach to predict the software faults," *SN Computer Science*, vol. 5, pp. 1-18, 2024, doi: https://doi.org/10.1007/s42979-024-02764-x.

[14] M. Magris and A. Iosifidis, "Bayesian learning for neural networks: an algorithmic survey," *Artificial Intelligence Review*, vol. 56, 11773–11823, 2023, doi: https://doi.org/10.1007/s10462-023-10443-1.

[15] A. F. Achmalia, Walid, and Sugiman, "Peramalan penjualan semen menggunakan backpropagation neural network dan recurrent neural network," *UNNES Journal of Mathematics*, vol. 9, no. 1, 2020, Available: https://journal.unnes.ac.id/sju/index.php/ujm/article/view/29970/16244.

[16] A. Bisry, C. M. S. Ramdani, and S. Yuliyanti, "Pengujian parameter algoritma genetika dan feed-forward neural networks pada permainan ular klasik," *MIND Journal*, vol. 9, no. 2, pp. 135–152, 2024, doi : https://doi.org/10.26760/mindjournal.v9i2.135-152.

[17] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun, "Hands-on bayesian neural networks-a tutorial for deep learning users," in *IEEE Computational Intelligence Magazine*, vol. 17, no. 2, pp. 29-48, 2020, doi: https://doi.org/10.1109/MCI.2022.3155327.

[18] A. Whata, K. Dibeco, K. Madzima, and I. Obagbuwa, "Uncertainty quantification in multi-class image classification using chest X-ray images of COVID-19 and pneumonia," *Frontiers in Artificial Intelligence*, vol. 7, 2024, doi: https://doi.org/10.3389/frai.2024.1410841.

[19] M. Conciatori, A. Valletta, and A. Segalini, "Improving the quality evaluation process of machine learning algorithms applied to landslide time series analysis," *Computers and Geosciences*, vol. 184, 2024, doi: https://doi.org/10.1016/j.cageo.2024.105531.

[20] B. Kocak, M. E. Klontzas, A. Stanzione, A. Meddeb, A. Demircioğlu, C. Bluethgen, K. K. Bressem, L. Ugga, N. Mercaldo, O. Díaz, and R. Cuocolo, "Evaluation metrics in medical imaging AI: fundamentals, pitfalls, misapplications, and recommendations," *European Journal of Radiology Artificial Intelligence*, 3, 100030, 2025, doi: https://doi.org/10.1016/j.ejrai.2025.100030.

[21] G. Zeng, "Invariance properties and evaluation metrics derived from the confusion matrix in multiclass classification," *Mathematics*, vol. 13, no. 16, 2025, doi: https://doi.org/10.3390/math13162609.

[22] S. Yang and G. Berdine, "Confusion matrix," *The Southwest Respiratory and Critical Care Chronicles*, vol. 12, no. 53, pp. 75–79, 2024, doi: https://doi.org/10.12746/swrccc.v12i53.1391.

https://doi.org/10.13057/ijas.v9i1.103994
E-ISSN: 2621-086X