Perbandingan Akurasi Model Pembelajaran Mesin SVM, KNN, *Decision Tree*, dan *Naive Bayes* pada Klasifikasi Gangguan Kesehatan Mental

Nurrahma[1], Tiya Muthia[2*], Yudi Eka Putra[3]

[1]Program Studi Teknik Informatika
Jl. Jalan Prof. Dr Jl. Prof. Dr. Ir. Sumantri Brojonegoro No.1, Bandar Lampung, Lampung 35141
[2,3]Program Studi Teknik Elektro
Jl. Jalan Prof. Dr Jl. Prof. Dr. Ir. Sumantri Brojonegoro No.1, Bandar Lampung, Lampung 35141
tiyamuthia@eng.unila.ac.id
nurrahma06@eng.unila.ac.id
yudiekaputra@mail.uml.ac.id

Intisari — **Gangguan kesehatan mental merupakan masalah yang terus meningkat di seluruh dunia, dengan kebutuhan mendesak akan diagnosis dini dan akurat. Pembelajaran mesin telah digunakan secara luas untuk mendukung proses diagnosis berbasis data. Penelitian ini membandingkan performa akurasi empat model pembelajaran mesin, yaitu *Support Vector Machine* (SVM), *K-Nearest Neighbor* (KNN), *Decision Tree* (DT), dan *Naive Bayes* (NB) dalam mengklasifikasikan gangguan kesehatan mental menggunakan dataset yang terdiri dari 17 fitur gejala dan satu label target dengan empat kelas diagnosis. Evaluasi dilakukan menggunakan metrik akurasi pada *test-set* (86,67%) dan *4-fold cross validation* (80.83%). Hasil menunjukkan bahwa model *Decision Tree* unggul dengan akurasi tertinggi, baik pada *test-set* maupun *4-fold cross validation*, mengindikasikan kemampuannya dalam menangani data heterogen dan memberikan interpretasi yang jelas. Penelitian ini memberikan wawasan tentang model pembelajaran mesin yang paling efektif untuk diterapkan dalam sistem diagnosis berbasis data di bidang kesehatan mental, yang berpotensi meningkatkan deteksi dini dan pengambilan keputusan klinis.**
Kata kunci — **Akurasi, *Cross Validation*, *Decision Tree*, KNN, *Naive Bayes*, SVM.**

Abstract — **Mental health disorders are a growing problem worldwide, with an urgent need for early and accurate diagnosis. Machine learning has been widely used to support data-driven diagnosis processes. This research compares the accuracy performance of four machine learning models, namely Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), and Naive Bayes (NB) in classifying mental health disorders using a dataset consisting of 17 symptom features and one target label with four diagnosis classes. Evaluation was carried out using accuracy metrics on the test-set (86.67%) and 4-fold cross validation (80.83%). The results show that the Decision Tree model is superior with the highest accuracy, both on the test-set and 4-fold cross validation, indicating its ability to handle heterogeneous data and provide clear interpretations. This research provides insight into the most effective machine learning models to apply in data-driven diagnosis systems in the mental health field, potentially improving early detection and clinical decision making.**
Keywords— **Accuracy, Cross Validation, Decision Tree, KNN, Naive Bayes, SVM**

I. PENDAHULUAN

Gangguan kesehatan mental merupakan salah satu isu kesehatan global yang signifikan dan terus meningkat. Menurut Organisasi Kesehatan Dunia (WHO), lebih dari 300 juta orang di seluruh dunia menderita depresi, menjadikannya salah satu penyebab utama disabilitas [1]. Selain itu, gangguan seperti kecemasan, gangguan bipolar, dan skizofrenia juga memberikan dampak yang besar terhadap individu dan masyarakat, termasuk peningkatan beban ekonomi dan sosial [2]. Diagnosis dini dan akurat menjadi penting untuk memastikan penanganan yang tepat, tetapi proses diagnosis berbasis gejala sering kali menghadapi tantangan karena kompleksitas dan subjektivitas penilaian klinis [3].

Dalam beberapa tahun terakhir, pembelajaran mesin (*machine learning*) telah digunakan secara luas dalam bidang kesehatan, termasuk untuk diagnosis gangguan kesehatan mental. Pembelajaran mesin memungkinkan analisis data besar dan kompleks secara cepat dan akurat, sehingga

dapat mengidentifikasi pola yang sulit dikenali oleh manusia . Model seperti *Support Vector Machine* (SVM), *K-Nearest Neighbor* (KNN), *Decision Tree* (DT), dan *Naive Bayes* (NB) telah banyak digunakan dalam klasifikasi data medis [4]. Masing-masing model memiliki kekuatan dan kelemahan tersendiri; misalnya, SVM efektif dalam ruang dimensi tinggi tetapi sensitif terhadap parameter, sementara KNN bergantung pada distribusi data di sekitar sampel tertentu. *Decision Tree* unggul dalam interpretasi dan menangani data heterogen, sedangkan *Naive Bayes* bekerja baik dalam data dengan asumsi independensi fitur [5].

Performa model pembelajaran mesin sangat dipengaruhi oleh karakteristik dataset, termasuk jumlah fitur, distribusi kelas, serta tingkat noise atau anomali dalam data [6]. Perbedaan ini memengaruhi efektivitas algoritma dalam mendeteksi pola dan membuat prediksi yang akurat. Oleh karena itu, membandingkan akurasi berbagai model menjadi penting untuk menentukan algoritma terbaik dalam konteks tertentu. Penelitian ini bertujuan untuk mengevaluasi performa SVM, KNN, Decision Tree, dan Naive Bayes dalam klasifikasi gangguan kesehatan mental menggunakan dataset dengan 17 fitur gejala dan satu label target dengan empat kelas diagnosis. Hasil penelitian diharapkan dapat memberikan wawasan tentang model yang paling sesuai untuk digunakan dalam mendukung sistem diagnosis berbasis data di bidang kesehatan mental.

II. METODOLOGI

Penelitian ini dilakukan melalui beberapa tahapan yang saling berkaitan, meliputi pengumpulan *dataset*, pemrosesan awal *dataset*, membangun model pembelajaran mesin, melatih model pembelajaran mesin, dan mengevaluasi model pembelajaran mesin.

A. Pengumpulan Dataset

Dataset yang digunakan dalam penelitian ini adalah *dataset* tentang pasien gangguan kesehatan mental yang didapatkan dari situs Kaggle. Dataset berupa *file* dengan format *Comma Separated Value* (CSV) yang berisi data terkait 120 pasien dengan 17 fitur berupa gejala penting dan 1 label diagnosis pakar yang dapat dilihat pada Tabel 1 dan Tabel 2. *Dataset* ini terakhir diperbarui pada 25 Januari 2024 [7] .

Tabel 1. Fitur Berupa Gejala Penting

No	Feature Name	Feature Type	Description
1	Sadness	Categorical	0: Seldom 1: Sometimes 2: Usually 3: Most-Often
2	Euphoric	Categorical	0: Seldom 1: Sometimes 2: Usually 3: Most-Often
3	Exhausted	Categorical	0: Seldom 1: Sometimes 2: Usually 3: Most-Often
4	Sleep disorder	Categorical	0: Seldom 1: Sometimes 2: Usually 3: Most-Often
5	Mood Swing	Boolean	0: No 1: Yes
6	Suicidal thoughts	Boolean	0: No 1: Yes
7	Anorexia	Boolean	0: No 1: Yes
8	Authority Respect	Boolean	0: No 1: Yes
9	Try-Explanation	Boolean	0: No 1: Yes
10	Aggressive Response	Boolean	0: No 1: Yes
11	Ignore & Move-On	Boolean	0: No 1: Yes
12	Nervous Break-down	Boolean	0: No 1: Yes
13	Admit Mistakes	Boolean	0: No 1: Yes
14	Overthinking	Boolean	0: No 1: Yes
15	Sexual Activity	Scale From 1 to 10	1-10
16	Concentration	Scale From 1 to 10	1-10
17	Optimism	Scale From 1 to 10	1-10

Tabel 2. Label Target Berupa Diagnosis Pakar

No	Label Name	Label Type	Description
1	Expert Diagnose	Multiclass	0: Normal 1: Bipolar Type-1 2: Bipolar Type-2 3: Depression

B. Pemrosesan Awal Dataset

Pemrosesan awal data adalah sebuah proses penting dalam pengembangan model pembelajaran mesin. Kumpulan data sering kali memiliki nilai yang hilang, nilai di luar rentang, dan lain sebagainya sehingga dapat mempengaruhi hasil akhir dari model pembelajaran mesin yang akan dikembangkan [8]. Dalam penelitian ini, fitur *dataset* merupakan data kategorikal dan tidak ada nilai yang hilang. Data kategorikal telah diberi label menggunakan fungsi `LabelEncoder` yang tersedia di *library* scikit-learn pada Python.

Dataset kemudian dibagi menjadi *training set* dan *testing set* secara acak menggunakan fungsi `train_test_split` yang tersedia di *library* scikit-learn pada Python dengan *`test size`*`=25%` dan *`random state`*`=0`.

C. Pemodelan Pembelajaran Mesin

Model pembelajaran mesin dibangun menggunakan *library* scikit-learn pada Python. Perubahan yang dilakukan pada algoritma dapat mempengaruhi hasil.

1) *Support Vector Machine (SVM)*: SVM adalah algoritma *supervised learning classification* yang diajukan oleh Vapnik pada tahun 1960-an. Secara geometris, pendekatan ini berfokus pada penentuan *hyperplane* optimal yang dapat memisahkan dua kelas atau klaster titik data dengan jarak yang sama dari keduanya. Awalnya, SVM dirancang untuk data yang terdistribusi secara linier. Namun, seiring waktu, fungsi *kernel* diperkenalkan untuk menangani kasus data yang tidak linier. [9]. Berikut adalah model SVM yang digunakan dalam penelitian ini:

```
svm_model = SVC()
svm_model.fit(X_train, y_train)
```

2) *K-Nearest Neighbor (KNN):* KNN adalah algoritma klasifikasi dasar yang non-parametrik. Algoritma KNN termasuk dalam *lazy learning algorithm*, tidak ada proses pembelajaran di sini. KNN tidak belajar berdasarkan data pelatihan, namun KNN melihat kelas tetangga terdekat yang bernomor *k* di seluruh kumpulan data. Nilai *k* dihitung terlebih dahulu dan mewakili jumlah elemen yang akan diperiksa. Tetangga data yang kelasnya harus ditentukan akan dianalisis dengan nilai *k* terdekat, dan dihitung jarak antar data tersebut. Dalam perhitungan jarak, KNN menggunakan jarak Euclidean [10]. Berikut adalah model KNN yang digunakan dalam penelitian ini:

```
knn_model =
KneighborsClassifier(n_neighbors =
5)
knn_model.fit(X_train, y_train)
```

3) *Decision Tree:* Salah satu metode *supervised learning classification* yang paling sukses adalah *decision tree* [9]. *Decision tree* sering digunakan untuk klasifikasi, pengelompokan, model prediksi, dan untuk membuat subkelompok dalam bidang penelitian terkait suatu masalah [11]. *Decision tree* embuat grafik atau pohon yang menggunakan teknik percabangan untuk menunjukkan setiap kemungkinan hasil suatu keputusan. Dalam representasi *decision tree*, setiap *node* internal menguji fitur, setiap cabang berhubungan dengan hasil dari *node* induk, dan setiap daun akhirnya menetapkan label kelas. Untuk mengklasifikasikan suatu *instance*, pendekatan *top-down* diterapkan mulai dari akar pohon. Untuk fitur atau *node* tertentu, cabang yang sesuai dengan nilai titik data untuk atribut tersebut dipertimbangkan hingga daun tercapai atau label ditentukan [9]. Pohon keputusan menggunakan regresi atau klasifikasi untuk memprediksi respons terhadap data. Regresi digunakan ketika data bersifat kontinu dan klasifikasi digunakan ketika fitur-fitur dikelompokkan. *Decision tree* dibuat dari simpul akar, cabang, dan simpul daun. Untuk mengevaluasi data, ikuti jalur dari simpul akar hingga mencapai simpul daun [12]. Berikut adalah model *decision tree* yang digunakan dalam penelitian ini:

```
dt_model =
DecisionTreeClassifier(max_depth =
5)
dt_model.fit(X_train, y_train)
```

4) *Naive Bayes (NB):* Metode *supervised learning classification* yang dikembangkan menggunakan Teorema probabilitas bersyarat Bayes dengan asumsi '*Naive'* bahwa setiap pasangan fitur saling independen. Artinya, dengan kata sederhana, kehadiran suatu fitur tidak dipengaruhi oleh kehadiran fitur lain dengan cara apa pun. Terlepas dari asumsi yang terlalu disederhanakan ini, pengklasifikasi NB berkinerja cukup baik dalam banyak situasi praktis. Hanya sejumlah kecil data pelatihan yang diperlukan untuk memperkirakan parameter tertentu [9]. Berikut adalah model *naive bayes* yang digunakan dalam penelitian ini:

```
nb_model = GaussianNB()
nb_model.fit(X_train, y_train)
```

D. K-fold Cross-Validation

K-fold Cross-Validation (K-fold CV) kemudian dilakukan menggunakan fungsi `cross_val_score` yang tersedia di *library* scikit-learn pada Python. *K-fold CV* merupakan teknik validasi dalam pembelajaran mesin yang digunakan untuk mengukur performa model secara lebih akurat, menghindari *overfitting*, dan memastikan bahwa model memiliki kemampuan generalisasi yang baik [6]. Dalam *K-fold CV*, *train-set* dibagi menjadi K *subset* (biasanya disebut "*fold*") dengan ukuran yang sama. Selama proses validasi, satu *fold* digunakan sebagai data validasi, sedangkan sisanya digunakan sebagai *train-set*. Proses ini diulang sebanyak K kali, sehingga setiap fold menjadi data validasi satu kali. Pada penelitian ini, *4-fold CV* dilakukan pada setiat model yang dibangun.

III. ANALISA DAN PEMBAHASAN

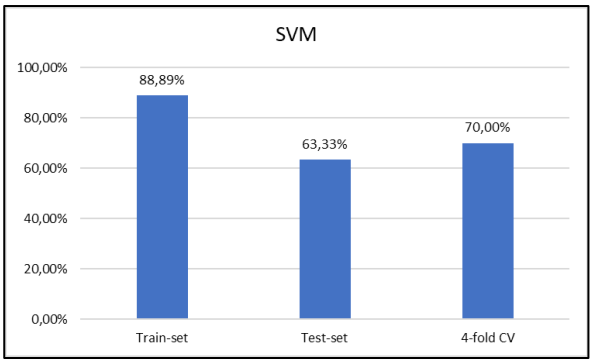
Pada penelitian ini, 4 (empat) model pembelajaran mesin dibangun untuk memprediksi kelas atau jenis gangguan kesehatan mental berdasarkan gejala-gejala yang ada. *Dataset* dibagi ke dalam *train-set* (75%) dan *test-set* (25%). *Train-set* digunakan untuk membangun model klasifikasi, sedangkan *test-set* digunakan untuk memvalidasi model yang dibangun. Selanjutnya masing-masing model divalidasi lebih lanjut menggunakan *4-fold CV.*

A. Hasil Model *Support Vector Machine (SVM)*

Hasil akurasi yang didapatkan oleh model SVM pada proses *training*, *testing*, dan *4-fold CV* dapat dilihat pada Tabel 3 dan Gambar 1.

Tabel 3. Hasil Akurasi Model SVM

SVM	**Akurasi**
Train-set	88.89%
Test-set	63.33%
4-fold CV	70.00%



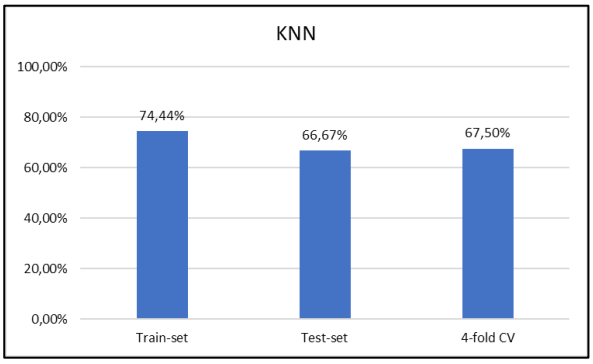
Gbr. 1 Grafik Hasil Akurasi Model SVM

B. Hasil Model *K-Nearest Neighbor (KNN)*

Hasil akurasi yang didapatkan oleh model KNN pada proses *training*, *testing*, dan *4-Fold CV* dapat dilihat pada Tabel 4 dan Gambar 2.

Tabel 4. Hasil Akurasi Model KNN

KNN	**Akurasi**
Train-set	74.44%
Test-set	66.67%
4-fold CV	67.50%



Gbr.2. Grafik Hasil Akurasi Model KNN

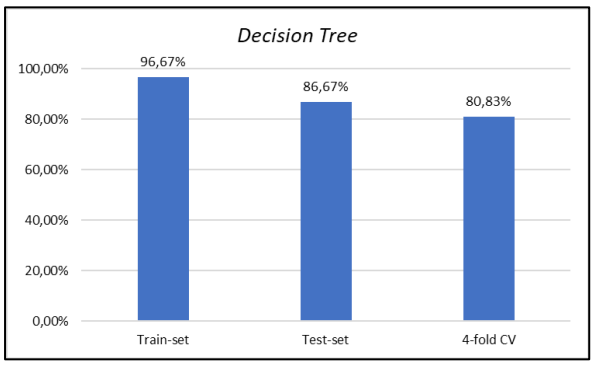
C. Hasil Model *Decision Tree*

Hasil akurasi yang didapatkan oleh model *decision tree* pada proses *training*, *testing*, dan

4-Fold CV dapat dilihat pada Tabel 5 Gambar 3.

Tabel 5. Hasil Akurasi Model *Decision Tree*

Decision Tree	**Akurasi**
Train-set	96.67%
Test-set	86.67%
4-fold CV	80.83%



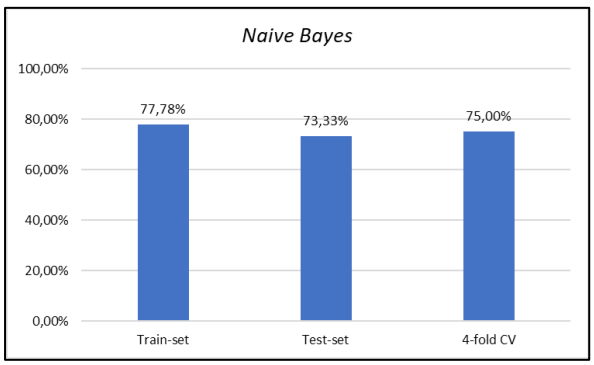
Gbr.3 Grafik Hasil Akurasi Model *Decision Tree*

D. Hasil Model *Naive Bayes*

Hasil akurasi yang didapatkan oleh model *naive bayes* pada proses *training*, *testing*, dan *4-Fold CV* dapat dilihat pada Tabel 6 dan Gambar 4.

Tabel 6. Hasil Akurasi Model *Naive Bayes*

Naive Bayes	**Akurasi**
Train-set	77.78%
Test-set	73.33%
4-fold CV	75.00%



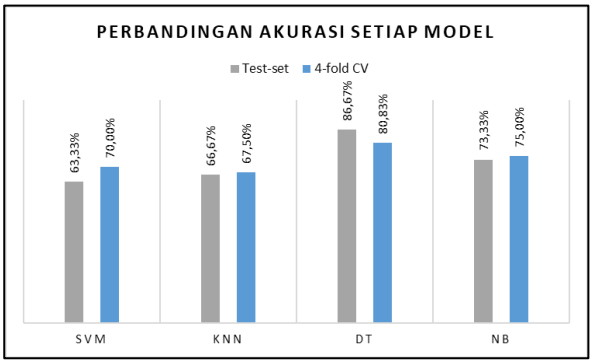
Gbr. 4. Grafik Hasil Akurasi Model *Naive Bayes*

E. Perbandingan Hasil Model Pembelajaran Mesin.

Perbandingan hasil akurasi yang didapatkan oleh setiap model pada proses *testing*, dan *4-Fold CV* dapat dilihat pada Tabel 7 dan Gambar 5.

Tabel 7. Perbandingan Hasil Akurasi Setiap Model

Akurasi	**SVM**	**KNN**	**DT**	**NB**
Test-set	63.33%	66.67%	**86.67%**	73.33%
4-fold CV	70.00%	67.50%	**80.83%**	75.00%



Gbr.5 Grafik Perbandingan Hasil Akurasi Setiap Model

Berdasarkan Tabel 7 dan Gambar 5, model *Decision Tree* (DT) menunjukkan akurasi tertinggi dibandingkan model lainnya, yaitu 86.67% pada *test set* dan 80.83% pada *4-fold CV*, mengindikasikan kemampuan generalisasi yang baik. KNN memiliki akurasi yang relatif stabil namun lebih rendah, yaitu 66,67% pada *test set* dan 67,50% pada *4-fold CV*. Naive Bayes (NB) dan SVM menunjukkan performa yang lebih baik daripada KNN, tetapi tetap di bawah DT, dengan NB sedikit lebih unggul daripada SVM. Hasil ini menunjukkan bahwa *Decision Tree* paling efektif dalam memodelkan dataset gangguan kesehatan mental ini, sementara KNN memiliki performa paling rendah.

IV. PENUTUP

Pada penelitian ini, model klasifikasi SVM, KNN, *Decision Tree*, dan *Naive Bayes* telah dibangun. Model-model klasifikasi tersebut diaplikasikan pada *dataset* pasien gangguan kesehatan mental yang didapatkan dari situs Kaggle. Dengan menggunakan *dataset* ini, *Decision Tree* (DT) memberikan performa terbaik untuk *test-set* (86,67%) dan *4-fold CV* (80,83%), menunjukkan bahwa model ini mampu menangkap pola dari *dataset* dengan baik dan memiliki kemampuan generalisasi yang baik. *Naive Bayes* dan SVM memiliki akurasi yang moderat, dengan NB lebih baik daripada SVM. KNN menunjukkan performa terendah, kemungkinan karena sensitivitas terhadap dimensi data atau distribusi sampel.

Untuk penelitian selanjutnya, disarankan melakukan *hyperparameter tuning* pada SVM dan KNN untuk menemukan konfigurasi optimal, seperti menyesuaikan parameter

kernel, *C*, dan *gamma* pada SVM, serta nilai *k* pada KNN. Selain itu, mempertimbangkan penggunaan *ensemble methods* seperti *Random Forest* atau *Gradient Boosting* dapat meningkatkan performa dengan menggabungkan kekuatan beberapa *Decision Trees*. Melakukan *feature selection* atau *scaling* juga dapat membantu model lain bekerja lebih optimal.

REFERENSI

[1] WHO, *Depression and Other Common Mental Disorders: Global Health Estimates*. Geneva: World Health Organization, 2017.

[2] T. Vos *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015," *Lancet*, vol. 388, no. 10053, pp. 1545–1602, 2016, doi: 10.1016/S0140-6736(16)31678-6.

[3] V. Patel *et al.*, "The Lancet Commission on global mental health and sustainable development," *Lancet*, vol. 392, no. 10157, pp. 1553–1598, Oct. 2018, doi: 10.1016/S0140-6736(18)31612-X.

[4] F. Jiang *et al.*, "Artificial intelligence in healthcare: Past, present and future," *Stroke Vasc. Neurol.*, vol. 2, no. 4, pp. 230–243, 2017, doi: 10.1136/svn-2017-000101.

[5] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier underZero-One Loss. Machine Learning," *Mach. Learn.*, vol. 29, pp. 103–130, 1997, [Online]. Available: https://link.springer.com/article/10.1023/A:1007413511361

[6] Pedro Domingos, "A Few Useful Things to Know About Machine Learning," *Commun. ACM*, vol. 55, no. 10, pp. 79–88, 2012, [Online]. Available: https://dl.acm.org/citation.cfm?id=2347755

[7] C. Desai, "Mental Disorder Classification." https://www.kaggle.com/datasets/cid007/mental-disorder-classification

[8] Nurrahma and R. Yusuf, "Comparing Different Supervised Machine Learning Accuracy on Analyzing COVID-19 Data using ANOVA Test," 2020. doi: 10.1109/ICIDM51048.2020.9339676.

[9] K. Das and R. N. Behera, "A Survey on Machine Learning : Concept , Algorithms and Applications," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 5, pp. 1301–1309, 2017, doi: 10.15680/IJIRCCE.2017.

[10] Rashmi Agrawal, "K-Nearest Neighbor for Uncertain Data," *Int. J. Comput. Appl.*, vol. 105, no. 11, pp. 13–16, 2014.

[11] Y. Ünal and M. N. Dudak, "Classification of Covid-19 Dataset with Some Machine Learning Methods," *J. Amasya Univ. Inst. Sci. Technol.*, 2020.

[12] T. Mythili, D. Mukherji, N. Padalia, and A. Naidu, "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)," *Int. J. Comput. Appl. Technol.*, vol. 68, pp. 10–15, 2013, doi: 10.5120/11662-7250.