Jurnal Komputer dan Informatika
Vol 17 No 1, April 2022: hlm 14-22

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

KLASIFIKASI TINGKAT KESELAMATAN PENYAKIT KANKER PAYUDARA MENGGUNAKAN METODE RANDOM FOREST

Wirya Aditya
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara
Jl. Letjen S.Parman No.1, Jakarta Barat, DKI Jakarta, Indonesia 11410
e-mail: *wirya.535200038@stu.untar.ac.id*

ABSTRAK

Penelitian ini mengkaji klasifikasi tingkat keselamatan penyakit kanker payudara menggunakan tiga metode berbeda: *Random Forest*, *Decision Tree*, dan *K-Nearest Neighbors* (KNN). Fokus utama adalah membandingkan hasil eksperimen dan mengevaluasi model dengan metrik evaluasi, *confusion matrix*, akurasi pelatihan dan pengujian, serta validasi silang. *Random Forest* secara konsisten mengungguli *Decision Tree* dan KNN dalam hal presisi dan akurasi karena keandalannya dalam menangani variasi data melalui beberapa pohon keputusan independen, menunjukkan keunggulan yang nyata meskipun perbedaan performa yang kecil. *Cross Validation* digunakan untuk memastikan bahwa model dapat menggeneralisasi dengan baik ke data yang belum pernah dilihat sebelumnya. Selain itu, pemantauan akurasi pelatihan dan pengujian membantu dalam mengevaluasi potensi *overfitting*. Penelitian ini menekankan pentingnya memilih metode klasifikasi yang sesuai dengan karakteristik unik dataset. Secara keseluruhan, temuan penelitian menyarankan bahwa *Random Forest* adalah pilihan optimal untuk klasifikasi tingkat keselamatan penyakit kanker payudara dalam dataset khusus ini. Namun, penting untuk selalu mempertimbangkan konteks dataset saat memilih metode klasifikasi yang paling sesuai untuk situasi tertentu. Hasil evaluasi menggunakan *cross validation* menunjukkan bahwa *Random Forest* merupakan indikator terbaik dengan tingkat akurasi yang paling tinggi, yaitu 91.49% dibandingkan *Decision Tree* dan *K-Nearest Neighbors*, yaitu 85.49 % and 90.42 %.

Kata kunci: *Random Forest, Decision Tree, K Nearest Neighbors,* Kanker Payudara, *Overfitting.*

ABSTRACT

This study explores the classification of breast cancer safety levels using three distinct methods: Random Forest, Decision Tree, and K-Nearest Neighbors (KNN). The primary focus is on comparing the experimental outcomes and evaluating the models using metrics evaluation, confusion matrices, train and test accuracies, and cross-validation. Random Forest consistently outperforms Decision Tree and KNN in terms of precision and accuracy due to its robustness in handling data variability through multiple independent decision trees, demonstrating a notable advantage despite small differences in performance. Cross-validation is employed to ensure the models generalize well to unseen data. Moreover, tracking train and test accuracies assists in assessing potential overfitting. The study underscores the importance of adapting the choice of classification method to the dataset's unique characteristics. In conclusion, the findings suggest that Random Forest is the optimal choice for breast cancer safety classification in this specific dataset. However, it is essential to consider dataset context when selecting the most suitable classification method for any given scenario. The evaluation results using cross-validation show that Random Forest is the best indicator with the highest accuracy rate, which is 91.49% compared to Decision Tree and K Nearest Neighbors, which are 85.49% and 90.42%.

Keywords: *Random Forest, Decision Tree, K Nearest Neighbors, Breast Cancer, Overfitting.*

1. PENDAHULUAN

Kanker payudara adalah salah satu jenis kanker yang paling umum dan menjadi salah satu penyebab utama kematian bagi perempuan di seluruh dunia. Penyakit ini adalah alasan pertama kematian karena pertumbuhan tumor ganas dalam waktu yang cukup lama, yaitu pada kisaran umur 20 dan 59 tahun dan yang kedua untuk wanita di atas umur 59 tahun. Diagnostik dini dan penilaian tingkat keselamatan penyakit menjadi krusial dalam manajemen penyakit ini. Selama beberapa dekade sebelumnya, penilaian konsisten yang diidentifikasi dengan investigasi pola tahap perkembangan ukuran tumor telah dilakukan. Mengingat besarnya pengobatan yang disesuaikan dan

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

pola penggunaan strategi *machine learning* yang berkembang, perkiraan dan tebakan terhadap keganasan tumor sudah dapat dilaksanakan [1] [2].

Identifikasi awal kanker payudara akan meningkatkan prognosis dan kemungkinan untuk bertahan hidup, karena hal ini akan meningkatkan proses penggandaan klinis yang tepat waktu kepada pasien. Selain itu prediksi akurat mengenai jenis dan ukuran tumor akan menurunkan tingkat pengobatan yang tidak penting. Dengan demikian, identifikasi kanker payudara yang tepat sesuai dengan tahap perkembangan tumor adalah subjek analisis yang cukup penting [3].

2. TINJAUAN LITERATUR

Beberapa klasifikasi untuk mengembangkan sistem pengklasifikasian dan prediksi penyakit kanker payudara. Pada penelitian ini, *Logistic Regression* (LR), *K-Nearest Neighbors* (KNN), dan *Support Vector Machine* (SVM) telah digunakan untuk melakukan klasifikasi. Hasil eksperimen menggunakan LR, KNN, dan SVM masing-masing akurasi 92,10%, 92,23%, dan 92,78% [4]. Dalam penelitian lain yang menggunakan metode yang sama, SVM menghasilkan akurasi sebesar 99,68% dan KNN menghasilkan akurasi sebesar 98,25% [5]. Berikut ini terdapat penelitian yang menggunakan dataset yang kurang baik, metode yang digunakan adalah KNN, *Decision Tree* (DT), *Naïve Bayes* (NB), dan *Artificial Neural Networks* (ANN). Eksperimen menunjukkan KNN menghasilkan akurasi 77,14%, DT menghasilkan akurasi 71,43%, NB menghasilkan akurasi 73,91%, dan ANN menghasilkan akurasi 80,0% [6].

Penelitian ini menggunakan semua metode yang ada sebelumnya dengan dataset yang cukup baik, Eksperimen menampilkan bahwa model DT, ANN, NB, LR, SVM, dan KNN masing-masing menghasilkan nilai akurasi 95,59%, 94,78%, 96,79%, 96,79%, 97,59%, dan 95,19% [7]. Model KNN dan LR yang menggunakan metode evaluasi *Confusion Matrix* dan *K-Fold Methods* masing-masing dapat dengan baik menghasilkan nilai akurasi sebesar 97,714% dan 96,52% [8]. Berikut terdapat model yang sama, yaitu LR, SVM, dan KNN tetapi menggunakan dataset yang berbeda. Hasil eksperimen menunjukkan LR menghasilkan akurasi 89,2%, SVM menghasilkan akurasi 91,6%, dan KNN menghasilkan 98,6% [9].

Penelitian berikut sudah melatih model KNN, NB, LR, dan SVM dan masing-masing model menghasilkan nilai akurasi yang akurat, yaitu 99%, 95%, 97%, dan 96% [10]. Pada penelitian ini, metode yang digunakan lebih detail, yaitu SVM-RBF-*Linear*, SVM-RBF-*Static*, SVM-RBF-*Dynamic*, LR *Generalized*, LR *Regularized*, KNN-*Euclidean*, KNN-*Manhattan*, NB normal, dan NB kernel. Masing-masing eksperimen menghasilkan nilai akurasi 80,58%, 64,03%, 93%, 90%, 92,08%, 95,68%, 94,96%, 92,1%, dan 92,1% [11]. Menggunakan dataset dari situs Kaggle dengan tingkat usability yang tinggi juga mempengaruhi tingkat akurasi metode *Decision Tree* untuk menghasilkan nilai akurasi 97,9021% [12].

Untuk penelitian in ada penambahan metode, yaitu *Ada Boost* dan *Binary Support Vector Machine* (BSVM) yang disertai metode KNN dan DT. Eksperimen masing-masing menghasilkan nilai akurasi 93,86%, 94,74% , 99,12%, dan 89,47% [13]. Tujuannya adalah untuk membangun beberapa model machine learning dari sekumpulan data pelatihan yang label kelas targetnya diketahui dan kemudian model ini digunakan untuk melakukan klasifikasi terhadap contoh yang tidak terlihat. Klasifikasi data kanker payudara akan berguna untuk memprediksi hasil dari beberapa penyakit atau menemukan perilaku genetik tumor [14].

3. METODE PENELITIAN

Metodologi dalam penelitian ini terdiri dari *Data collection*, *Data pre-processing*, *Model development*, dan *Model Evaluation*.

3.1 *Data Collection*

Dataset diunduh dari situs *Kaggle*. Data set terdiri dari 4024 baris dan 16 kolom, yaitu 15 fitur dan 1 target (*Age, Race, Marital Status, T Stage, N Stage, 6th Stage, Differentiate, Grade, A Stage, Tumor Size, Estrogen Status, Progesterone Status, Regional Node Examined, Regional Node Positive, Survival Month, Status*).

3.2 *Data Pre-processing*

Untuk memastikan integritas data yang digunakan dalam analisis, beberapa pra-pemrosesan data telah dilakukan dalam metodologi penelitian ini. Pertama, proses pengecekan untuk mengetahui apakah ada nilai null (*missing value*) dalam kumpulan data. Kedua, pengecekan terhadap data duplikat dalam kumpulan data. Data duplikat dapat mempengaruhi hasil model, sehingga prosedur yang digunakan untuk mengidentifikasi dan menghapus data duplikat telah dilakukan. Ketiga, analisis data *outlier* juga dilakukan. *Outlier* adalah data yang berbeda secara signifikan dari dataset lainnya dan dapat mempengaruhi hasil analisis statistik. Terakhir, *label encoder* diberikan pada variabel kategorikal untuk menyiapkan data sebelum digunakan dalam model klasifikasi. Hal ini penting karena model klasifikasi memerlukan data dalam bentuk numerik, sehingga *label encoder* diperlukan untuk mengubah variabel kategorikal menjadi bentuk numerik yang sesuai untuk model.

3.3 *Model Development*

Dalam model development, dataset akan dipisah menjadi 2 bagian, yaitu fitur (x) dan target (y). kemudian, dataset akan dibagi menjadi dua, yaitu data latihan (*train data*) dan data uji (*test data*) untuk menguji model. Metode klasifikasi yang akan diterapkan ada 3 macam, yaitu *Random Forest*, *Decision Tree*, dan *K-Nearest Neighbors* (KNN). Setiap model akan dilatih dengan data latihan dan dievaluasi dengan data uji.

a. *Random Forest*

Random Forest adalah algoritma *supervised learning* yang mudah untuk beradaptasi dan digunakan. *Random Forest* seperti namanya *Forest* atau "hutan" terbuat dari beberapa *decision tree* dan kemudian menggabungkan mereka untuk mendapatkan satu yang akurat. Cara kerja *Random Forest* terbagi menjadi dua bagian. Bagian pertama adalah menggabungkan hasil dari beberapa *decision tree* lalu membentuk sebuah hutan. Bagian kedua adalah membuat prediksi dari setiap *decision tree* yang sudah dibuat pada bagian pertama [15].

Berikut ini adalah langkah-langkahnya:

i. Algoritma akan memilih sampel data secara acak dari dataset yang sudah disediakan.
ii. Algoritma membuat *decision tree* untuk setiap sampel data yang dipilih. Kemudian hasil prediksi dari setiap *decision tree* akan didapatkan.
iii. Terjadinya proses *voting* setiap hasil prediksi. Khusus untuk klasifikasi nilai yang akan dipakai adalah nilai *modus* (nilai yang paling sering muncul).
iv. Hasil akhir adalah hasil prediksi yang memiliki nilai *voting* paling banyak.

b. *Decision Tree*

Decision Tree adalah algoritma klasifikasi dengan bentuk seperti pohon yang memiliki akar, ranting, dan daun. Akar (*internal node*) akan mewakili fitur dataset. Ranting (*branch node*) akan mewakili aturan keputusan. Daun (*leaf node*) akan mewakili hasil. Pembuatan *decision tree* akan terbalik dari atas ke bawah yang diawali dengan akar dan turun sampai daun. Dalam pemilihan keputusan pada *decision tree*, akan terbentuk *subtree* pilihan untuk setiap kemungkinan hasil.

Decision tree menjadi salah satu algoritma paling efisien dan terbaik untuk klasifikasi dan prediksi. Pada Gambar 1 dibawah ini menampilkan ilustrasi dari cara kerja *decision tree* [16].

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

Gambar 1. Contoh Cara Kerja *Decision Tree*

c. *K-Nearest Neighbors* (KNN)

K-Nearest Neighbors (KNN) adalah algoritma yang bekerja berdasarkan kemiripan objek dan objek tersebut cenderung berada pada jarak yang dekat antara satu dengan yang lain. Oleh karena itu, data yang serupa atau memiliki karakteristik yang cukup sama, cenderung saling bertetangga. KNN melakukan klasifikasi terhadap *instance* suatu objek berdasarkan mayoritas tetangganya. "*K*" adalah bilangan bulat yang selalu positif. Tetangga dipilih dari sekumpulan objek yang tidak memiliki klasifikasi yang jelas. Ada atribut numerik "*N*" dimensi yang digunakan untuk menjelaskan sampel pelatihan.

Semua sampel pelatihan disimpan dalam ruang pola "*N*" dimensi karena setiap sampel mewakili titik dalam ruang berdimensi "*N*". Klasifikasi KNN mencari pola untuk "*K*" sampel pelatihan yang paling dekat dengan sampel yang tidak diketahui. "Kedekatan" didefinisikan dengan jarak Euclidean. Sampel yang tidak diketahui paling sering diberi kelas di antara "*K"* tetangga terdekatnya. Kelas sampel pelatihan yang paling dekat dengan sampel yang tidak diketahui diberikan dalam ruang pola ketika K=1 [17].

4. HASIL DAN PEMBAHASAN

Dalam pembahasan hasil penelitian ini, model akan dievaluasi dengan menggunakan 4 macam metode, yaitu *Cross-Validation*, *Evaluation Metrics*, *Confusion Matrix*, dan Train and Test *Accuracy*. Hasil percobaan akan ditampilkan dalam bentuk gambar ataupun tabel. Tujuan dari melakukan *cross validation* adalah untuk menghitung rata-rata akurasi setiap model [18].

Untuk hasil perhitungan, *Random Forest* memiliki nilai rata-rata akurasi sebesar 91,49%, *Decision Tree* memiliki nilai rata-rata akurasi sebesar 85,49%, dan KNN memiliki nilai rata-rata akurasi sebesar 90,42%. Setiap nilai tersebut dapat di lihat pada Gambar 2 untuk hasil perhitungan akurasi dengan metode *cross validation* dan Gambar 3 menampilkan grafik hasil *cross validation* untuk ketiga model. Dapat dilihat bahwa, model random forest memiliki rata-rata yang lebih tinggi dibandingkan model lainnya.

Evaluation Metrics digunakan untuk mengetahui nilai *precision*, *recall*, dan *f1-score*. Nilai *precision* adalah nilai hitung rasio antara jumlah prediksi positif dibagi dengan total prediksi positif. Nilai *recall* adalah nilai sensitivitas untuk mengukur seberapa jauh model benar dalam melakukan prediksi kepada kasus positif. Nilai *f1-score* adalah nilai rata-rata antara nilai *precision* dan *recall*.

```
cv_random_forest = cross_val_score(random_forest_classifier, x, y, cv=5) * 100
print("Random Forest:")
print(f"Cross-Validation Mean Accuracy: {cv_random_forest.mean():.2f}%")
```

```
Random Forest:
Cross-Validation Mean Accuracy: 91.49%
```

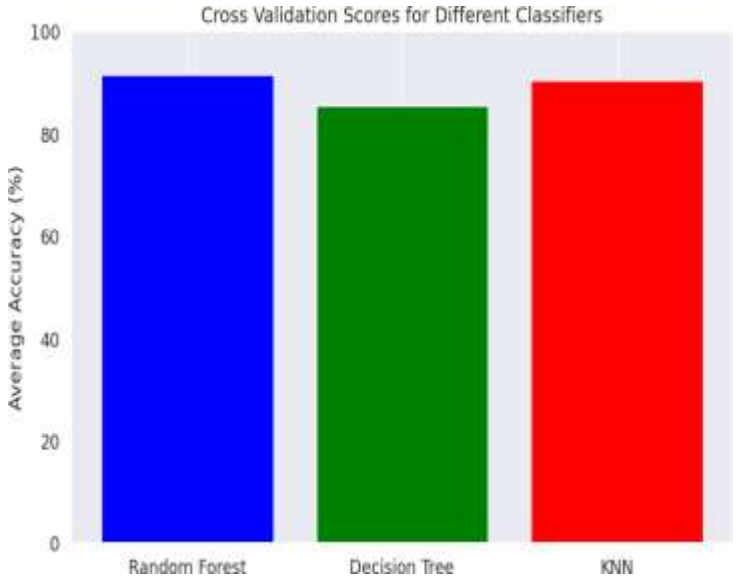
```
cv_decision_tree = cross_val_score(decision_tree_classifier, x, y, cv=5) * 100
print("Decision Tree:")
print(f"Cross-Validation Mean Accuracy: {cv_decision_tree.mean():.2f}%")
```

```
Decision Tree:
Cross-Validation Mean Accuracy: 85.49%
```

```
cv_knn = cross_val_score(knn_classifier, x, y, cv=5) * 100
print("K Nearest neighbors:")
print(f"Cross-Validation Mean Accuracy: {cv_knn.mean():.2f}%")
```

```
K Nearest neighbors:
Cross-Validation Mean Accuracy: 90.42%
```

Gambar 3. Hasil Perhitungan *Cross Validation*



Gambar 4. Grafik Hasil *Cross Validation*

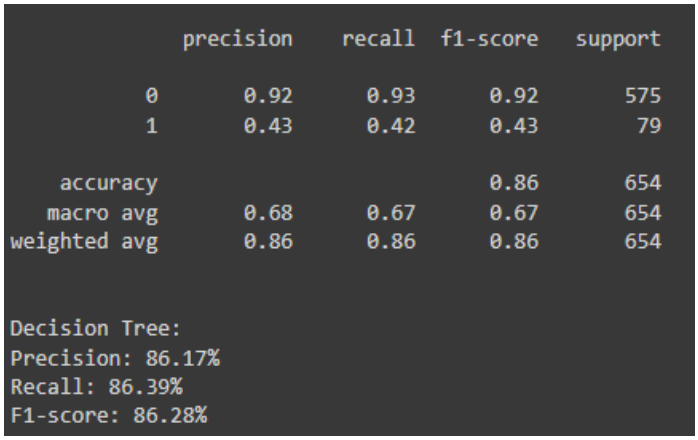
Pada Gambar 5 ditunjukkan laporan *evaluation metrics* untuk model *random forest* yang dapat dilihat bahwa *random forest* memiliki nilai *precision* sebesar 90,54%, nilai *recall* sebesar 91,28%, dan nilai *f1-score* sebesar 89,81%.

```
              precision    recall  f1-score   support

           0       0.92      0.99      0.95       575
           1       0.81      0.37      0.50        79

    accuracy                           0.91       654
   macro avg       0.86      0.68      0.73       654
weighted avg       0.91      0.91      0.90       654

Random Forest:
Precision: 90.54%
Recall: 91.28%
F1-score: 89.81%
```

Gambar 5. Hasil Perhitungan *Evaluation Metrics* untuk *Random Forest*

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

Pada Gambar 6 berikut ini ditunjukkan laporan *evaluation metrics* untuk model *decision tree*. Dapat dilihat bahwa *decision tree* memiliki nilai *precision* sebesar 86,17%, nilai *recall* sebesar 86,39%, dan nilai *f1-score* sebesar 86,28%.

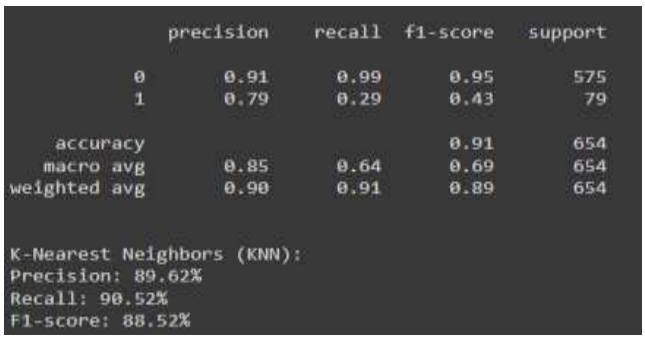


	precision	recall	f1-score	support
0	0.92	0.93	0.92	575
1	0.43	0.42	0.43	79
accuracy			0.86	654
macro avg	0.68	0.67	0.67	654
weighted avg	0.86	0.86	0.86	654

Decision Tree:
Precision: 86.17%
Recall: 86.39%
F1-score: 86.28%

Gambar 5. Hasil Perhitungan *Evaluation Metrics untuk Decision Tree*

Pada Gambar 7 dibawah ini menampilkan laporan *evaluation metrics* untuk model *k-nearest neighbors*. Dapat dilihat bahwa *k-nearest neighbors* memiliki nilai *precision* sebesar 89,62%, nilai *recall* sebesar 90,52%, dan nilai *f1-score* sebesar 88,52%.



	precision	recall	f1-score	support
0	0.91	0.99	0.95	575
1	0.79	0.29	0.43	79
accuracy			0.91	654
macro avg	0.85	0.64	0.69	654
weighted avg	0.90	0.91	0.89	654

K-Nearest Neighbors (KNN):
Precision: 89.62%
Recall: 90.52%
F1-score: 88.52%

Gambar 7. Hasil Perhitungan *Evaluation Metrics* untuk *K-Nearest Neighbors*

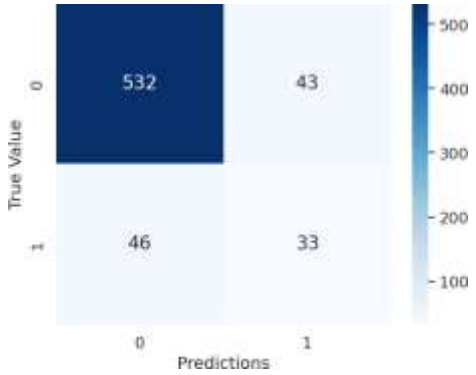
Berikutnya akan dilakukan evaluasi menggunakan metode *confusion matrix*. *Confusion Matrix* adalah alat untuk menampilkan prediksi model benar atau salah. Di dalam metode evaluasi *confusion matrix* terdapat empat nilai yang perlu diperhatikan, yaitu *True Positive*, *True Negative*, *False Positive*, dan *False Negative*.

True Positive (TP) adalah nilai matriks untuk menghitung berapa kali model memprediksi kelas positif dan nilai sebenarnya juga kelas positif. *True Negative* (TN) adalah nilai matriks untuk menghitung berapa kali model memprediksi kelas negatif dan nilai sebenarnya juga kelas negatif. *False Positive* (FP) adalah nilai matriks untuk menghitung berapa kali model memprediksi kelas positif dan nilai sebenarnya adalah kelas negatif. *False Negative* (FN) adalah nilai matriks untuk menghitung berapa kali model memprediksi kelas negatif dan nilai sebenarnya adalah kelas positif [19].

Berikut ini pada Gambar 8 menampilkan total hasil yang diperoleh dengan metode evaluasi *confusion matrix* untuk model *random forest*. Terdapat sebanyak 568 kali memprediksi status hidup dan sebenarnya status juga hidup, sebanyak 29 kali memprediksi status meninggal dan sebenarnya juga meninggal, sebanyak 50 kali memprediksi status hidup tetapi sebenarnya status meninggal, sebanyak 7 kali memprediksi status meninggal tetapi sebenarnya status hidup.

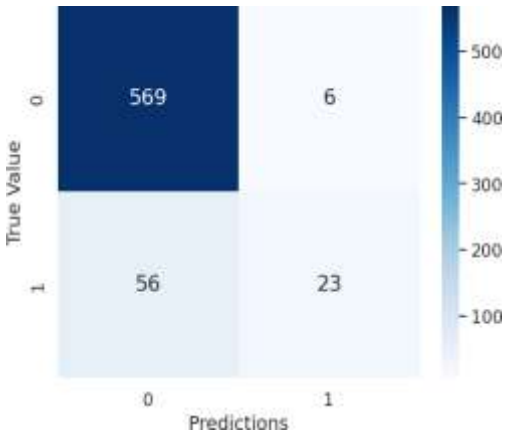
Gambar 8. Hasil Evaluasi *Confusion Matrix* untuk *Random Forest*

Pada Gambar 9 dibawah ini akan ditampilkan menampilkan total hasil yang diperoleh dengan metode evaluasi *confusion matrix* untuk model *decision tree*. Terdapat sebanyak 532 kali memprediksi status hidup dan sebenarnya status juga hidup, sebanyak 33 kali memprediksi status meninggal dan sebenarnya juga meninggal, sebanyak 46 kali memprediksi status hidup tetapi sebenarnya status meninggal, sebanyak 43 kali memprediksi status meninggal tetapi sebenarnya status hidup.



Gambar 9. Hasil Evaluasi *Confusion Matrix* untuk *Decision Tree*

Pada Gambar 10 dibawah ini akan ditampilkan menampilkan total hasil yang diperoleh dengan metode evaluasi *confusion matrix* untuk model *k-nearest neighbors*. Dapat diketahui bahwa sebanyak 569 kali memprediksi status hidup dan sebenarnya status juga hidup, sebanyak 23 kali memprediksi status meninggal dan sebenarnya juga meninggal, sebanyak 56 kali memprediksi status hidup tetapi sebenarnya status meninggal, sebanyak 6 kali memprediksi status meninggal tetapi sebenarnya status hidup.



Gambar 10. Hasil Evaluasi *Confusion Matrix* untuk *K-Nearest Neighbors*

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

Terakhir akan dilakukan evaluasi menggunakan *train and test accuracy*, metode ini digunakan untuk mengetahui apakah model mengalami *overfitting* atau *underfitting*. *Overfitting* dan *Underfitting* akan menunjukkan pengaruh data baru yang ingin digunakan untuk memprediksi terhadap efisiensi model, jika terjadi *overfitting* maka akurasi data latih akan lebih besar daripada data uji, jika terjadi *underfitting* maka akurasi data latih dan data uji sama-sama rendah. Hal ini akan mempengaruhi seberapa baik model dapat memprediksi data baru untuk diprediksi [20]. Berikut pada Tabel 1 berikut ini menunjukkan hasil *train* dan *test* untuk masing-masing model.

Tabel 1. Hasil *Train* dan *Test* Akurasi Ketiga Model

Model	Train accuracy	Test accuracy
RF	100%	91,28%
DT	100%	86,39%
KNN	92,23%	90,52%

Dapat ditemukan bahwa ketiga model mulai dari *random forest*, *decision tree*, dan *k-nearest neighbors* dapat memprediksi data baru yang masuk dengan baik. Model *random forest* adalah model terbaik untuk memprediksi target dengan nilai rata-rata akurasi yang tinggi, nilai error yang rendah, dan termasuk ke dalam kategori *good fit*.

5. KESIMPULAN

Dalam penelitian ini, model *random forest* memiliki nilai akurasi paling tinggi di antara model *decision tree* dan *k-nearest neighbors*. Perlu dilakukan analisis lebih lanjut dan *hyperparameter tuning* terhadap model agar hasil klasifikasi status hidup atau meninggal menjadi lebih efisien dan maksimal. Pada saat penelitian berlangsung ketiga model tidak mengalami *overfitting* dan menghasilkan nilai rata-rata akurasi 91,49%.

DAFTAR PUSAKA

[1] T. a. F. G. Kadir, "Lung cancer prediction using machine learning and advanced imaging techniques," Translational lung cancer research, vol. III, no. 7, 2018.
[2] Y. Xiao, "A deep learning-based multimodel ensemble method for cancer prediction.," Computer methods and programs in biomedicine, p. 153, 2018.
[3] B. S. Ma, "02 Brain cancer prediction using machine learning methods and high-throughput molecular data.," 2017.
[4] K. P. S. S. Ch. Shravya, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. VIII, no. 6, pp. 1106-1110, 2019.
[5] H. I. M. R. H. a. M. K. H. M. M. Islam, "Prediction of breast cancer using support vector machine and K-Nearest neighbors," in IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, Bangladesh, 2017.
[6] T. M. A. a. F. H. J. M. U. Ghani, "Comparison of Classification Models for Early Prediction of Breast Cancer," in 2019 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, 2019.
[7] D. G. A. R. d. K. G. Ravi Kumar, "An Efficient Prediction of Breast Cancer Data using Data Mining Techniques," International Journal of Innovations in Engineering and Technology (IJIET), vol. II, no. 4, 2013.
[8] V. K. V. M. A. A. Y. a. A. J. T. Jain, "Supervised Machine Learning Approach For The Prediction of Breast Cancer," in 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 2020.
[9] V. S. Madhu Kumari, "Breast Cancer Prediction system," Procedia Computer Science, vol. 132, pp. 371-376, 2018.
[10] G. Singh, "Breast Cancer Prediction Using Machine Learning," International Journal of Scientific Research in Computer Science, Engineering and Information Technology | IJSRCSEIT, vol. VI, no. 4, pp. 278-284, 2020.

[11] P. C. A. D. N. K. Mandeep Rana, "Breast Cancer Diagnosis And Recurrence Predictionusing Machine Learning Techniques," International Journal of Research in Engineering and Technology | IJRET, vol. IV, no. 4, pp. 372-376, 2015.
[12] M. O. M. H. H. A. M. T. M. A. A. Omar Tarawneh, "Breast Cancer Classification using Decision Tree Algorithms," International Journal of Advanced Computer Science and Applications | IJACSA, vol. XIII, no. 4, 2022.
[13] B. C. A. T. O. D. a. S. H. S. Laghmati, "Classification of Patients with Breast Cancer using Neighbourhood Component Analysis and Supervised Machine Learning Techniques," in 2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet), Marrakech, Morocco, 2020.
[14] A. Z. K. A. B. H. G. D. S. M. Y. Q. H. L. a. B. Z. Morteza Heidari, "Prediction of breast cancer risk using a machine learning approach embedded with a locality preserving projection algorithm," Institute of Physics and Engineering in Medicine, vol. 63, no. 3, 2018.
[15] L. W. F. X. X. &. Z. S. Lin, "Random forests-based extreme learning machine ensemble for multi-regime time series prediction," Expert Systems with Applications, vol. 83, p. 164–176.
[16] B. M. K. a. S. A. S. Murugan, "Classification and Prediction of Breast Cancer using Linear Regression, Decision Tree and Random Forest," in Mysore, IndiaInternational Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), Mysore, India, 2017.
[17] A. R. a. S. A. D. Cahyanti, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," Indones. J. Data Sci., vol. I, no. 2, pp. 39-43, 2020.
[18] D. S. P. a. D. S. B. P. I. Nainggolan, "Klasifikasi Informasi Kesehatan Pada Data Media Sosial Menggunakan Support Vector Machine dan K-Fold Cross Validation," Malikussaleh J. Mech. Sci. Technol., vol. V, no. 2, pp. 34-38, 2021.
[19] A. J. H. H. Ikhsan Nuh Atthalla, "Klasifikasi Penyakit Kanker Payudara Menggunakan Metode K Nearest Neighbor (KNN)," Annual Research Seminar | ARS, vol. IV, no. 1, pp. 148-151, 2018.
[20] W. S. J. S. A. R. &. A. A. Z. Saputra, "Seleksi Fitur Menggunakan Random Forest Dan Neural Network," Islamic Education Studies | IES, pp. 978-979, 2011.