e-ISSN 2962-1968

Diagnosis Penyakit Hepatitis Menggunakan Fuzzy K-NN dan Ensemble Learning

Ery Permana Yudha[1*], Eko Purwanto[2], Joni Maulindar[3]

[1,2,3]*Sistem Informasi*
Universitas Duta Bangsa Surakarta
[1*]*ery_permanayudha@udb.ac.id*, [2]*eko_purwanto@udb.ac.id*, [3]*joni_maulindar@udb.ac.id*

Abstrak— **Penyakit hepatitis adalah satu diantara lima penyakit yang paling mematikan. Hepatitis menyerang organ hati yang diakibatkan oleh serangan virus. Penyakit hepatitis merupakan masalah kesehatan global yang mempengaruhi jutaan orang di seluruh dunia. Deteksi dini dan pengklasifikasian yang akurat dari penyakit hepatitis sangat penting untuk pengelolaan dan perawatan yang tepat. Dalam artikel ini, kami mengusulkan sebuah sistem deteksi penyakit hepatitis yang menggunakan kombinasi metode Fuzzy k-NN dan Ensemble Learning. Pendekatan ini bertujuan untuk meningkatkan akurasi dan kehandalan dalam diagnosis hepatitis. Metode klasifikasi yang tepat dan data informasi yang lengkap dapat membantu dalam mendeteksi kemungkinan harapan hidup seseorang yang terkena penyakit hepatitis. Oleh karena itu kami mengusulkan untuk menggunakan gabungan metode yang merupakan gabungan dari Fuzzy K-NN dengan *ensemble learning*. *Ensemble learning* yang digunakan pada penelitian ini adalah *ensemble bagging* dan *adaptive boosting*. *Ensemble learning* digunakan untuk meningkatkan akurasi dari Fuzzy K-NN untuk memprediksi kemungkinan harapan hidup seseorang yang terkena penyakit hepatitis. Sebelum dilakukan metode klasifikasi dilakukan preprocessing terhadap dataset hepatitis dengan mengisi *missing value* dan menyeimbangkan data dengan menggunakan metode *Synthetic Minority Oversampling Technique* (SMOTE) *borderline*. Hasil uji coba pada penelitian ini menunjukkan bahwa metode yang diusulkan mendapatkan rata-rata akurasi sebesar 94,87% dan melebihi kinerja dari metode lain yang digunakan sebagai bahan uji.**
Kata kunci— Hepatitis, fuzzy k-nn, ensemble learning, bagging, boosting.

Abstract— **Hepatitis is one of the five most deadly diseases. Hepatitis attacks the liver organ and is caused by viral infections. Hepatitis is a global health problem that affects millions of people worldwide. Early detection and accurate classification of hepatitis are crucial for proper management and treatment. In this article, we propose a hepatitis detection system that combines Fuzzy k-NN and Ensemble Learning methods. This approach aims to improve accuracy and reliability in hepatitis diagnosis. Accurate classification methods and comprehensive data can help detect the life expectancy of individuals with hepatitis. Therefore, we propose using a combination of Fuzzy k-NN and ensemble learning methods. The ensemble learning techniques used in this research are ensemble bagging and adaptive boosting. Ensemble learning is employed to enhance the accuracy of Fuzzy k-NN in predicting the life expectancy of individuals with hepatitis. Before the classification method, preprocessing is performed on the hepatitis dataset by handling missing values and balancing the data using the Synthetic Minority Oversampling Technique (SMOTE) borderline. The experimental results of this study show that the proposed method achieves an average accuracy of 94.87% and outperforms other methods used as benchmarks.**
Keywords— Hepatitis, fuzzy k-nn, ensemble learning, bagging, boosting.

I. PENDAHULUAN

Penyakit hepatitis merupakan salah satu penyakit yang menyerang organ hati akibat dari serangan infeksi virus. gejala utama penyakit hepatits misalnya muntah, demam, urin bewarna gelap, nyeri lambung, berat badan turun, dan lain-lain. Berdasarkan data dari World Health Organization (WHO) ada 325 juta penduduk dunia terinfeksi hepatitis B dan C dan 1,4 juta orang meninggal dunia tiap tahun karenanya. Sehingga hepatitis merupakan salah satu penyakit mematikan di dunia. Penyakit hepatitis memiliki beberapa tingkatan kerusakan hati mulai dari yang terendah yaitu munculnya fibrosis pada hati hingga yang paling berbahaya yang dapat merusak jaringan parut pada hati yang dapat menyebabkan kematian. Beberapa penelitian telah dilakukan untuk memprediksi tingkat hepatitis dengan mengukur jumlah fibrosis pada hati penderita penyakit. Selain itu ada juga penelitian dengan cara mengukur jumlah resistensi insulin untuk memprediksi kadar fibrosis yang terdapat pada hati penderita hepatitis C.

Penelitian tentang hepatitis sudah banyak dilakukan oleh peneliti dengan berbagai macam fokus salah satunya yaitu dari mendeteksi tingkat kerusakan hati, jumlah fibrosis, dan mendeteksi kemungkinan hidup seorang penderita penyakit hepatitis. Beberapa peneliti dalam mendeteksi kemungkinan hidup penderita penyakit diperlukan beberapa parameter yang diambil seperti usia, jenis kelamin, dan indikator lain seperti steroid, ukuran liver, kadar albumin, dan lain sebagainya. Selain parameter diperlukan juga sebuah metode klasifikasi yang baik dan kelengkapan informasi pada dataset

agar menghasilkan ketepatan prediksi yang tinggi. Jika banyak data yang missing value atau kosong maka tidak bisa menghasilkan sebuah model klasifikasi yang tepat. Sehingga akan menghasilkan nilai akurasi yang rendah.

Penelitian ini kami menggunakan dataset dari UCI Machine Learning dengan jumlah fitur sebanyak 20 dan 155 data. Namun, dataset tersebut masih terdapat beberapa data yang kosong bahkan perbandingan jumlah kelasnya juga tidak seimbang atau disebut dengan class imbalance. Sehingga diperlukan suatu cara untuk menangani permasalahan tersebut. Ada beberapa penelitian yang membahas cara mengatasi class imbalance dan missing value. Setelah permasalahan dataset diselesaikan selanjutnya menentukan sebuah model klasifikasi yang baik agar bisa memprediksi kelas dengan baik. Pada penelitian ini kami mengusulkan sebuah metode klasifikasi dengan menggabungkan Fuzzy k-Nearest Neighbor dengan Ensemble Learning yang diterapkan pada data yang kurang berkualitas. Selain itu penelitian ini kami juga menggunakan metode khusus untuk permasalahan ketimpangan kelas pada dataset.

Selanjutnya pada penulisan paper untuk penelitian ini disusun sebagai berikut. Bab II berisi tentang tinjauan umum tentang penelitian yang terkait. Bab III merupakan penjelasan secara rinci tentang metode penelitian yang diusulkan. Kemudian pada bab IV dijelaskan hasil pengujian yang telah dilakukan dengan menggunakan metode yang telah diusulkan. Dan yang terakhir bab V merupakan kesimpulan dari penelitian yang dikerjakan serta saran apa yang harus dilakukan untuk penelitian selanjutnnya.

II. METODOLOGI PENELITIAN

Pembelajaran mesin dapat digunakan untuk mengklasifikasi data dari berberapa fitur yang tersedia. Bidang kesehatan merupakan area dalam pembelajaran mesin karena memiliki berbagai sumber data yang sulit jika dilakukan pengklasifikasian secara manual. Hepatitis merupakan salah satu dari 5 besar penyakit yang mematikan yang ada di dunia [1]. Dataset penyakit hepatitis yang terdapat dalam UCI masih tidak seimbang antara jumlah data dengan kelas sakit dan jumlah data dengan kelas sehat. Kelas sakit berjumlah 32 data dan kelas sehat berjumlah 123. Untuk mengatasi data yang tidak seimbang ini maka diperlukan suatu algoritma yang efektif. Synthetic minority oversampling technique (SMOTE) merupakan salah satu algoritma data sampling yang efektif dan telah diterapkan untuk menganalisa data medis yang masih tidak seimbang [2].

Dalam klasifikasi diperlukan metode klasifikasi yang baik agar mendapatkan hasil yang bagus. Fuzzy k-nearest neighbor (FKNN) merupakan salah satu metode klasifikasi pengembangan dari metode k-nearest neighbor dengan menggabungkan fuzzy set theory [3]. Dalam paper tersebut dengan menggunakan metode klasifikasi FKNN didapatkan hasil akurasi yang cukup memuaskan. Dari berberapa penelitian yang telah dilakukan metode ensemble dapat meningkatkan akurasi dari klasifikasi [4]. Pada penelitian ini kami akan menggunakan SMOTE untuk mengatasi ketidak seimbangan data, FKNN sebagai metode classifier, dan meningkatkan akurasi dari FKNN dengan menggunakan metode ensemble. Maka dari itu, kami melihat penelitian yang terkait secara terpisah.

D. *Synthetic Minority Oversampling Technique (SMOTE)*

Pada penelitian ini dataset yang digunakan terdapat ketidakseimbangan. Jika kumpulan data yang ditujukan untuk klasifikasi memiliki jumlah data yang tidak sama untuk kelas yang berbeda, maka kumpulan data tersebut dikatakan menjadi tidak seimbang. Pelatihan model pembelajaran mesin dengan dataset yang tidak seimbang ini, sering menyebabkan model menimbulkan bias tertentu terhadap kelas mayoritas. Untuk mengatasi masalah ini telah dilakukan banyak penelitian. Salah satunya dengan metode oversampling, yaitu dimana kita menambah kelas minor agar kelas minor sama banyaknya dengan kelas mayor.

Salah satu metode oversampling adalah synthetic minority oversampling technique (SMOTE). SMOTE sebelumnya terbukti telah mengurangi ketimpangan kelas pada studi kasus medis [5-8]. Penelitian borderline SMOTE pernah dilakukan pada kasus metastatis otak dari kanker paru-paru [5]. SMOTE dapat mengatasi persoalan seperti deteksi fraud pada transaksi kartu kredit [9], deteksi spam [10], dan prediksi software yang cacat atau tidak

sempurna [11]. Dari berberapa penelitian tersebut terbukti bahwa SMOTE dapat mengatasi ketidak seimbangan data.

E. *Fuzzy k-Nearest Neighbor (FKNN)*

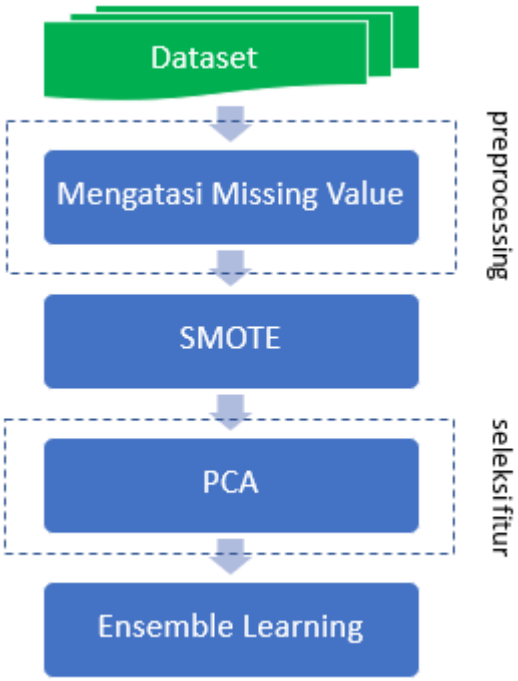
Algoritma fuzzy k-nearest neighbor (FKNN) adalah metode klasifikasi yang memasukkan fuzzy set theory ke dalam KNN. Sehingga, dalam perancangannya tidak memerlukan persamaan matematis yang kompleks dari objek yang akan dikendalikan. Penelitian sebelumnya dilakukan pada studi kasus klasifikasi gender dengan menggunakan levehnstein based fuzzy knn [12], prediksi tinggi gelombang pada danau [13] dan prediksi pada penyakit parkinson [14]. Pada masing-masing penelitian tersebut dapat disimpulkan metode FKNN berhasil mengatasi permasalahan dengan baik. Dan pada ketiga penelitian tersebut FKNN mendapatkan hasil akurasi yang baik sehingga kami termotivasi untuk menggunakan metode tersebut dalam penelitian ini.

F. *Ensemble Learning*

Metode ensemble adalah suatu algoritma pada pembelajaran mesin atau machine learning yang menggunakan beberapa algoritma learning lainnya untuk mendapatkan solusi prediksi yang lebih baik daripada algoritma tanpa ensemble. Ensemble adalah algoritma yang sebelumnya telah terbukti meningkatkan akurasi pada Naïve Bayes, Bayes Net, C4.5, dan MLP [1]. Penelitian sebelumnya juga membahas penggabungan antara Ensemble dan SMOTE yaitu SMOTE-boost [5] dan SMOTE-bagging [6-8] yang dimuat pada paper tahun 2020[4]. Pada semua penelitian yang telah dilakukan terbukti ensemble learning dapat meningkatkan kinerja dari suatu algoritma.

III. HASIL DAN PEMBAHASAN

Pada tahap ini kami akan menjelaskan tahapan-tahapan yang dilakukan pada penelitian ini. Proses pada penelitian ini secara garis besar terdiri dari beberapa tahapan yaitu mengatasi dataset yang mengandung missing value. Melakukan oversampling menggunakan SMOTE pada dataset dikarenakan dataset hepatitis tidak seimbang. Selanjutnya dilakukan seleksi fitur menggunakan PCA dengan tujuan untuk memilih fitur terbaik dari suatu kumpulan data fitur, sehingga tidak mengurangi kompleksitas perhitungan. Dan yang terakhir dilakukan klasifikasi menggunakan ensemble learning yang digabungkan dengan fuzzy knn (FKNN). Secara umum alur metode pada penelitian ini ditunjukkan pada Gambar 1.



Gambar 18. Alur metode penelitian

G. *Preprocessing*

Dataset hepatitis yang didapatkan dari UCI machine learning repository masih ada beberapa data yang tidak lengkap. Dikatakan tidak lengkap karena data masih ada yang mengandung missing value. Ketidaklengkapan data dapat menyebabkan misclassification. Maka untuk mengatasi masalah tersebut perlu dilakukan preprocessing. Untuk menangani missing value, preprocessing yang dilakukan adalah dengan mengisi nilai pada cell kosong berdasarkan rata-rata jika atributnya adalah numerik, dengan mempertimbangkan kelas yang sama pada data.

Namun ketika atributnya adalah kategorikal, untuk mengisi *missing value* tersebut kita menggunakan modus dengan mempertimbangkan kelas yang sama pada data. Dengan begitu data yang tadinya kosong akan terisi dengan data baru berdasarkan mean atau modus dari masing-masing fitur untuk setiap kelas.

H. *Borderline SMOTE*

Selain itu jumlah data diantara kedua kelas pada dataset hepatitis masih tidak seimbang. Kemudian untuk menangani ketidakseimbangan data, pada kelas minor dilakukan oversampling pada kelas minor dengan menggunakan metode borderline-SMOTE. Borderline SMOTE merupakan pengembangan dari algoritma oversampling SMOTE.

Berbeda dengan algoritma SMOTE yang tidak mempertimbangkan letak dari kelas minor terhadap kelas major, borderline SMOTE mempertimbangkan letak dari kelas minor terhadap kelas major. Pada borderline SMOTE kelas minor akan dioversampling jika jumlah kelas major pada nearest neighbor kelas minor memenuhi persamaan. Jika tidak memenuhi persamaan maka kelas minor tidak dioversampling.

1. **For** each X_i in S_{min} **do**
2. **Set** m-Nearest Neighbor of X_i
3. **If** $(m/2) \leq | S_{j\ m\text{-}NN} \cap S_{major} | < m$ **Then**
4. **Set** û secara random ambil $S_{j\ m\text{-}NN}$
5. $X_{new} = X_i + (û\text{-}X_i)\ \delta$
6. **EndIf**
7. **EndFor**

Gambar 19. Pseudocode

Pada Gambar 2 dijelaskan bahwa X_i adalah X pada indeks ke-I dari S_{min}. Dimana S_{min} adalah himpunan dari kelas minor. $S_{i:m\text{-}NN}$ adalah himpunan tetangga dari X_i. S_{major} adalah himpunan kelas major. û adalah X tetangga dari kelas X_i. δ adalah variabel distribusi random. $S_{i:m\text{-}NN}$ merupakan nearest neighbor dari kelas minor.

I. *Seleksi Fitur*

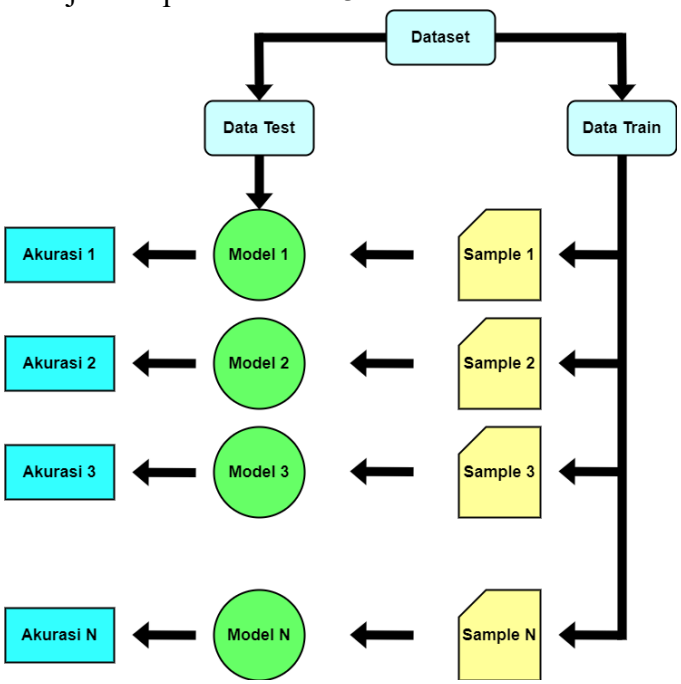
Seleksi fitur adalah teknik untuk memilih fitur penting dan relevan terhadap data dan mengurangi fitur yang tidak relevan. Seleksi fitur bertujuan untuk memilih fitur terbaik dari suatu kumpulan data fitur. Seleksi fitur dilakukan menggunakan PCA dengan dilakukan pencarian pada eigenvector tertinggi berdasarkan nilai eigenvalue. Kemudian dilakukan transformasi dari dimensi asli ke dimensi reduksi. Berikut adalah langkah-langkah PCA:

1. Buat matrix O dimana kolom adalah atribut.
2. Kurangi rata-rata tiap kolom matrix O hingga menjadi nol. kemudian masukkan ke matrix M.
3. Buatlah Single Value Decomposition M dimana $M=U\Sigma V^T$. U adalah matrix orthonormal berisi Left Singular Vector, Σ adalah matrix diagonal dan V adalah matrix eigenvektor kanan dari M.
4. Pilih reduksi dimensi sebanyak L dan pilih hanya singular vektor sebanyak L pertama dari U_L untuk mereduksi M. Kemudian buat matrix baru $N=(U_L)^T M$.

J. *Ensemble Learning*

Ensemble adalah strategi yang dapat digunakan untuk meningkatkan akurasi dari algoritma klasifikasi. Ini adalah teknik klasifikasi meta yang efektif yang menggabungkan model pembelajar yang lemah dengan model pembelajar yang kuat untuk meningkatkan akurasi dari model pembelajar yang lemah.

Dalam penelitian ini, teknik ensemble digunakan untuk meningkatkan akurasi algoritma FKNN untuk prediksi penyakit hepatitis. Tujuan menggabungkan beberapa pengklasifikasi adalah untuk memperoleh kinerja yang lebih baik dibandingkan dengan pengklasifikasi individual. Prosedur untuk ensemble ditunjukkan pada Gambar 3.



Gambar 20. Tahapan ensemble learning

Ensemble learning yang digunakan pada penelitian ini yaitu boosting dan baging. Pada boosting data training dibagi menjadi beberapa subhimpunan. Kemudian classifier dilatih dengan

e-ISSN 2962-1968

subset untuk menghasilkan model. Apabila ada elemen terjadi misklasifikasi maka elemen itu dimasukkan ke dalam subhimpunan yang baru.

Sedangkan, bagging dikenal dengan nama bootstrap aggregation. Bagging memilih pola secara random dari data latih dengan pengembalian. Data latih yang baru dibuat memiliki jumlah pola yang sama dengan data latih asli dengan beberapa pengecualian dan pengulangan. Data latih yang baru dikenal dengan replikasi bootstrap. Pada bagging sampel bootstrap diambil dari data dan classifier dilatih dengan setiap sampel. Voting dari setiap classifier dikumpulkan dan hasil dari classifier dihitung berdasarkan majority vote.

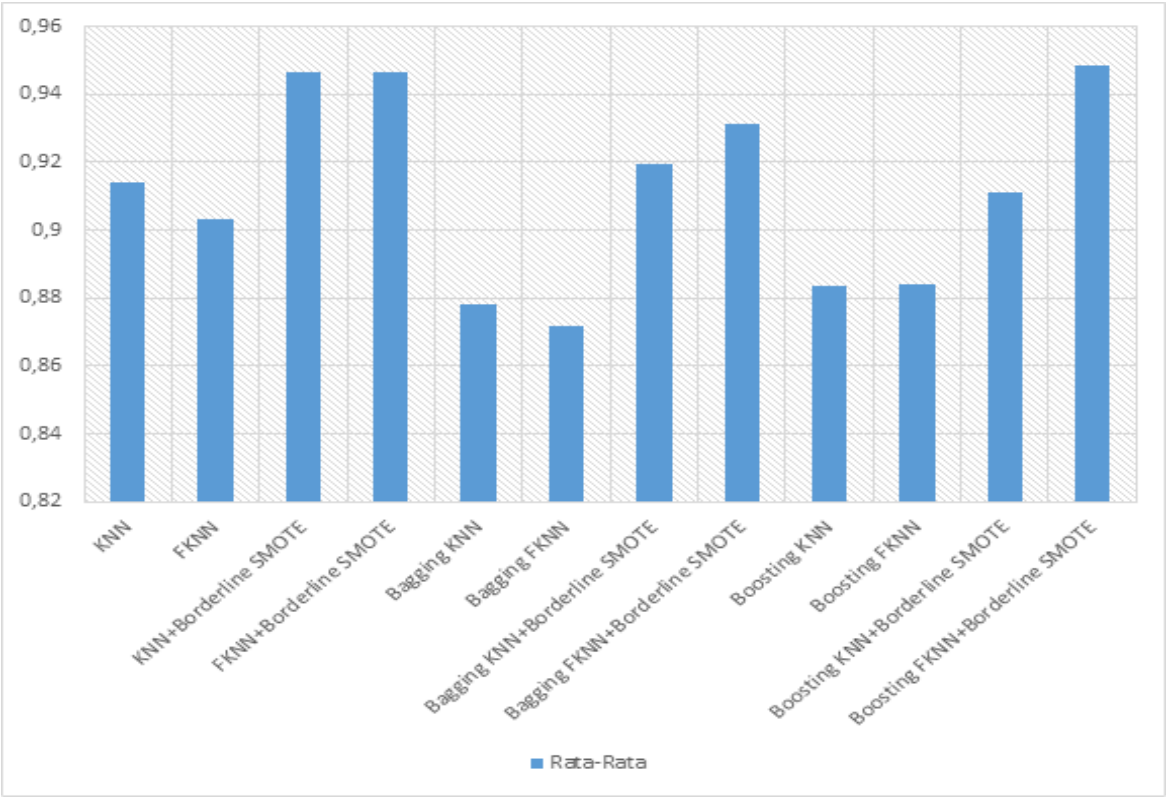
K. Pengujian

Dalam penelitian ini kami menggunakan dataset hepatitis. Dataset hepatitis terdiri dari 19 Atribut yang terdiri dari 12 data kategorikal dan 7 data numerik. Responden terdiri dari laki-laki dan perempuan berumur 20-78 yang didapatkan informasi dari tenaga medis dan bukan dari pasien.

Ada empat skenario pengujian yang kami lakukan yaitu melakukan klasifikasi menggunakan FKNN dan KNN tanpa borderline-SMOTE dan tanpa ensemble, klasifikasi dengan borderline-SMOTE tanpa ensemble, klasifikasi tanpa borderline-SMOTE dengan ensemble, dan klasifikasi dengan borderline-SMOTE dengan ensemble. Ensemble learning yang kita gunakan hanya dua yaitu bagging dan boosting.

Pada Gambar 4 menunjukkan grafik hasil pengujian dari keempat skenario menggunakan FKNN dan KNN. Grafik menunjukkan bahwa metode oversampling borderline-SMOTE dapat meningkatkan akurasi dari pengklasifikasian. Grafik juga menunjukkan boosting FKNN dengan borderline-SMOTE menghasilkan akurasi paling tinggi dibandingkan metode yang lain dengan akurasi sebesar 94,87%. Sementara hasil rata-rata akurasi terendah didapatkan ketika melakukan klasifikasi bagging FKNN tanpa borderline-SMOTE yaitu sebesar 87,2%.

IV. KESIMPULAN



Gambar 21. Hasil rata-rata akurasi setiap skenario pengujian

Jumlah data pada dataset hepatitis sebanyak 155 data.

Penelitian ini digunakan untuk mengenalkan penggabungan FKNN dengan teknik ensemble pada studi kasus penyakit hepatitis. Studi ini menggunakan dataset hepatitis yang didapatkan dari UCI machine learning repositori digunakan sebagai data latih dan data testing. Ensemble bagging dan ensemble boosting digunakan pada penelitian. Sebagai data pembanding kita menggunakan klasifikasi lain yaitu KNN. Dari skenario pengujian tersebut didapatkan hasil yang terbaik yaitu skenario penggunaan borderline-SMOTE dengan gabungan FKNN dan ensemble boosting dengan akurasi rata-rata sebesar 94,87%. Dari hasil uji coba tersebut, metode fuzzy k-nn yang diusulkan mampu meningkatkan akurasi yang dikombinasikan dengan algoritma borderline SMOTE.

REFERENSI

[1] M. Rouhani and M. M. Haghighi, "The diagnosis of hepatitis diseases by support vector machines and artificial neural networks," *2009 Int. Assoc. Comput. Sci. Inf. Technol. - Spring Conf. IACSIT-SC 2009*, pp. 456–458, 2009.

[2] M. Zeng, B. Zou, F. Wei, X. Liu, and L. Wang, "Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data," *Proc. 2016 IEEE Int. Conf. Online Anal. Comput. Sci. ICOACS 2016*, vol. 2016, pp. 225–228, 2016.

[3] H. L. Chen *et al.*, "An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach," *Expert Syst. Appl.*, vol. 40, no. 1, pp. 263–271, 2013.

[4] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics Med. Unlocked*, vol. 16, no. June, p. 100203, 2019.

[5] H. Ghaderi Zefrehi and H. Altınçay, "Imbalance learning using heterogeneous ensembles," *Expert Syst. Appl.*, vol. 142, 2020.

[6] Chawla, N. V. , Lazarevic, A. , Hall, L. O. , & Bowyer, K. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. In Knowledge discovery in databases: PKDD 2003 (pp. 107–119). Springer Berlin Heidelberg .

[7] Lertampaiporn, S. , Thammarongtham, C. , Nukoolkit, C. , Kaewkamnerdpong, B. , & Ru- engjitchatchawalya, M. (2013). Heterogeneous ensemble approach with discrim- inative features and modified-smotebagging for pre-mirna classification. Nucleic Acids Research, 41 .

[8] Galar, M. , Fernandez, A. , Barrenechea, E. , Bustince, H. , & Herrer, F. (2012). A review on ensembles for the class imbalance problem: bagging, boosting, and hybrid-based approaches. IEEE Transactions on Systems Man and Cybernetics Part C, 42 , 463–484 .

[9] Wang, S. , & Yao, X. (2009). Diversity analysis on imbalanced data sets by using ensemble models. In IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009 (pp. 324–331) .

[10] Panigrahi, S., Kundu, A., Sural, S., & Majumdar, A. (2009). Credit card fraud detection: a fusion approach using dempster-shafer theory and bayesian learning. Information Fusion, 10 , 354–363. doi: 10.1016/j.inffus.20 08.04.0 01 .

[11] Lu, X.-Y. , Chen, M.-S. , Wu, J.-L. , Chang, P.-C. , & Chen, Z. (2018). A novel ensemble decision tree based on under-sampling and clonal selection for web spam detection. Pattern Analysis and Applications, 21 , 741–754 .

[12] Wang, S. , & Yao, X. (2013). Using class imbalance learning for software defect prediction. IEEE Transactions on Reliability, 62 , 434–443 .

[13] O. Dogan and B. Oztaysi, "Genders prediction from indoor customer paths by Levenshtein-based fuzzy kNN," *Expert Syst. Appl.*, vol. 136, pp. 42–49, 2019

[14] M. R. Nikoo, R. Kerachian, and M. R. Alizadeh, "A fuzzy KNN-based model for significant wave height prediction in large lakes," *Oceanologia*, vol. 60, no. 2, pp. 153–168, 2018.