KNOWCON
Knowledge on Economics and Management
International Scientific Conference
Recent Advances in Business Analytics. Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics pp. 29–35
DOI: 10.15439/2021B5
ISSN 2300-5963 ACSIS, Vol. 29

Using the generalized fuzzy k-nearest neighbor classifier for biomass feedstocks classification

Mahinda Mailagaha Kumbure, Pasi Luukka
LUT University
Yliopistonkatu 34, 53850 Lappeenranta, Finland
Email: {mahinda.mailagaha.kumbure, pasi.luukka}@lut.fi

***Abstract*—This paper proposes a novel framework based on a recently introduced classifier called multi-local power mean fuzzy k-nearest neighbor (MLPM-FKNN) and the Minkowski distance to classify biomass feedstocks into property-based classes. The proposed approach uses k nearest neighbors from each class to compute class-representative multi-local power mean vectors and the Minkowski distance instead of the Euclidean distance to fit the most suitable distance metric based on the properties of the data in finding the nearest neighbors to the new data point. We evaluate the performance of the proposed approach using three biomass datasets collected from several articles published in reputable journals and the Phyllis 2 biomass database. Input features of the biomass samples include their characteristics from the proximate analysis and ultimate analysis. In the developed framework, we interpret the biomass feedstocks classification as a five-class problem, and the classification performance of the proposed approach is benchmarked with the results obtained from classical k-nearest neighbor-, fuzzy k-nearest neighbor- and support vector machine classifiers. Experimental results show that the proposed approach outperforms the benchmarks and verify its effectiveness as a suitable tool for biomass classification problems. It is also evident from the results that the features from both ultimate and proximate analyses can offer a better classification of biomass feedstocks than the features considered from each of those analyses separately.**

***Index Terms*—Biomass feedstocks, Fuzzy k-nearest neighbor, Machine learning, Minkowski distance, Proximate properties, Ultimate properties**

I. Introduction

BIOMASS is a biological material obtained from living organisms such as animals and plants. Biomass feedstocks are diverse, usually derived from agricultural residues, forest products waste, food waste, green waste, municipal solid waste, and other waste [1]. Due to its organic nature and abundant supply, biomass is considered as an essential renewable energy source [2] and has received much attention in the world [1]. Biomass is typically used to derive various energy products, for example, biogas, bioethanol, biodiesel, and solid fuel [3]. Following oil, coal, and natural gas, biomass has been the fourth largest energy source globally to date [4].

Primary concerns regarding biomass investigations include enhancing and extending the general understanding of the biomass properties and compositions, and also using this knowledge for achieving sustainable development in energy generation [5]. In the study of biomass, in general, two different types of analyses: proximate analysis and ultimate analysis, are used to determine the nature of biomass in terms of the chemical compounds [6]. The proximate analysis is applied to measure the compositions of volatile matter, moisture, ash and fixed carbon in the biomass. On the basis of ash and moisture content, ultimate analysis yields the amount of carbon (C), hydrogen (H), nitrogen (N), oxygen (O), and sulfur (S) [6]. These properties and their classification corresponding to the various biomass materials are considered more important when they are selected as energy feedstocks [5]. The energy conversion process has also encouraged the studies for biomass feedstocks classification considering their properties such as proximate properties, thermal properties, chemical properties, to mention few [7].

Artificial intelligence, particularly machine learning (ML), has been extensively used to analyze various types of data classification and prediction problems effectively. However, applying ML-based techniques in biomass analysis is still a new development [8]. In the literature, a few studies have focused on the potential of some ML techniques for biomass classification and related research. Tao et al. [9] used a principal component analysis (PCA) based approach to attribute the biomass properties within five groups. Wang et al. [10] also applied the PCA to find the most influential features of biomass for the decision-making process in bioenergy production. Olatunji et al. [5] attempted to grade the biomass feedstocks based on their proximate properties using k-nearest neighbor (KNN) method. The best performance they found with the KNN model [11] was around 70% in the training and validation. A recent study by [8] examined the effectiveness of several ML techniques, including Random Forest, KNN, Gaussian Naïve Bayes, and Decision Tree models to predict and differentiate biomass types based on the Pyrolysis molecular beam mass spectrometry (py-MBMS) analyses. They showed that the KNN classifier generally performed the best compared to others. The present work introduces a novel ML-based approach for biomass classification by interpreting the classification task as a five-class problem.

Our proposed approach is based on the multi-local power mean fuzzy k-nearest neighbor (MLPM-FKNN) method that is an enhanced version of the KNN classifier, which was recently introduced in [12]. This new KNN method is chosen as it has showed more robust to outliers and random variables than original ones according to [12]. This technique can perform well in situations where clear imbalances in class distributions of the data are found [12]. In this study, we generalize the

performance of the MLPM-FKNN classifier using k nearest neighbors from each class to compute class-representative multi-local power mean vectors. In addition to that, we also introduce the Minkowski distance for the k nearest neighbor search in the learning part instead of the Euclidean distance to fit the most suitable distance metric according to the data properties in finding the nearest neighbors for the unclassified data point from each class. Since the Minkowski distance is a generalized distance of the Euclidean and Manhattan distances, its utilization allows greater flexibility for obtaining more relevant neighboring points close to the unclassified data point.

To examine the classification performance of the proposed approach, we use three biomass datasets collected from several articles [7], [13], [14], [15], [16] and the Phyllis 2 biomass database [17]. Four well-known performance measures are used to assess the performance of the proposed method, and the observed results are benchmarked with three state-of-art techniques such as the KNN, fuzzy k-nearest neighbor (FKNN) [18], and support vector machine (SVM) [19] classifiers. From the wide variety of machine learning techniques [20], [21], these were chosen since they are similar to proposed method and easily available. In summary, the main contributions of this paper include (i) proposing a generalized MLPM-FKNN classifier with Minkowski distance for biomass classification, (ii) using chemical compound features derived from ultimate analysis for biomass classification, and empirically examining whether they have a great influence on the classification of biomass, (iii) applying biomass data from Phyllis 2 data repository for classification purpose, and (iv) comparing the classification performance of the proposed intelligent model with the performance of several well-known ML techniques.

II. Preliminaries

This section briefly presents the preliminaries of relevant k-nearest neighbor classifier variants, the Power mean operator, and the Minkowski distance measure. In addition, the Minkowski distance-based generalized MLPM-FKNN classifier is introduced.

A. KNN and FKNN Classifiers

The KNN classifier [11] is a simple, effective, and robust supervised machine learning technique. Due to its high accuracy and capability in the pattern classification, the KNN classifier has been widely used in many real-world applications (for examples, see [22], [23]). It begins with calculating the Euclidean distances from the query sample (i.e., unclassified data point) to the training instances. Then, a set of k nearest neighbors is identified for the query sample from the sorted training instances in ascending order according to the Euclidean distances measured. Finally, the query sample is assigned to the class represented by the majority of the nearest neighbors. However, the KNN method intuitively suffers from some weaknesses. For instance, it gives equal importance to all nearest neighbors neglecting the fact that different instances have different impacts on the classification of the query sample [24]. Moreover, it does not take into account the strength of the class membership for the query sample [25]. To deal with these issues, the FKNN model [18] has been introduced as an enhancement of the original algorithm.

In the FKNN, the set of k nearest neighbors of the query sample (Q) is searched first as in the KNN classifier. After that, a membership degree for each class is measured for the query sample using weighted distances from k nearest neighbors to the query sample. Lastly, it classifies the query sample into the class with the highest membership degree among all classes. To compute the class memberships (u_i for class i) for Q, the formula used can be defined as follows:

$$u_i(Q) = \frac{\sum_{j=1}^{k} u_{ij}(1/\|Q - X_j\|^{2/(r-1)})}{\sum_{j=1}^{k}(1/\|Q - X_j\|^{2/(r-1)})} \tag{1}$$

where, $r \in (1, +\infty)$ is a fuzzy strength parameter and u_{ij} is the membership degree of the j^{th} nearest neighbor X_j from the i^{th} class. Also,

To compute u_{ij}, there are two main approaches: one is through the crisp membership, and the other is through the fuzzy membership [18]. In this study, we use the crisp labeling approach where the full membership is assigned to the known class and zero memberships to all other classes.

B. Power Mean and Minkowski Distance

Power mean (also called generalized mean) is a function of means. If $\{x_1, x_2, .., x_m\}$ is a set of real numbers and p is a real parameter, then power mean (M_p) is defined as:

$$M_p = (\frac{1}{m}\sum_{l=1}^{m} x_l^p)^{1/p} \text{ for } p \neq 0 \tag{2}$$

When $p \to 0$, $M_p \to \prod_{i=1}^{m} X_i^{1/m}$. With the power mean function, different types of means can be generated including well-known harmonic mean ($p = -1$), arithmetic mean ($p = 1$), and quadratic mean ($p = 2$). Additionally, M_p approaches to geometric mean when $p \to 0$.

The Minkowski distance (also referred to as L_p norm) between two data points $X = \{x_1, x_2, .., x_m\}$ and $Y = \{y_1, y_2, .., y_m\}$ in m-dimensional space is defined as follows:

$$d_{Md}(X, Y) = \big(\sum_{l=1}^{m} |x_l - y_l|^q\big)^{1/q} \text{ for } q \geq 1 \tag{3}$$

The Minkowski distance represents a class of distance functions that are formed by the parameter q. For instance, by setting $q = 1$, we obtain the Manhattan distance (also called City block distance). Similarly, the Euclidean distance is observed in the case of $q = 2$.

C. Modified MLPM-FKNN Classifier

The concept of the multi-local power mean fuzzy k-nearest neighbor (MLPM-FKNN) classifier is easy to understand. It has been developed by introducing a local-mean computation into the learning part of the FKNN method. The local mean

vectors are calculated for each class in the set of nearest neighbors by using the power mean function. These vectors are called multi-local power mean vectors. In this way, the MLPM-FKNN method creates "representative vectors" for each class to perceive the class information for query sample instead of comparing it directly to the k-nearest neighbors. Also, changing the power mean parameter allows us to find its best possible options, which will enhance the classification accuracy [12].

In this study, we deploy a generalized version of this method. The Minkowski distance function is applied according to the study by [5] instead of the Euclidean distance to measure the distances from the query sample to the training instances. The purpose of using Minkowski distance here is to generate greater flexibility for obtaining more relevant neighbors close to the query sample since it has an optimizable parameter to adjust the function to the data set available. A formal definition of the developed method can be presented as follows.

Let $\{X_j, c_j\}_{j=1}^{n}$ be a training set with n instances, where $X_j = \{x_j^1, x_j^2, .., x_j^m\}$ is an input instance j from m-dimensional feature space, and its output class label is $c_j \in C$ ($C = \{\omega_1, \omega_2, .., \omega_T\}$: the set of class labels and T is the number of classes). For a given query sample $Q = \{q^1, q^2, .., q^m\}$, the goal is to fit the classifier from the training set in order to predict the class ω^* for Q. The steps of the generalized MLPM-FKNN classifier in this study can be presented as follows:

(i) Group the training data $\{X_j, c_j\}_{j=1}^{n}$ into each class ω_i. The resulting class subsets can be denoted as $\{X_j, \omega_i\}_{j=1}^{n_i}$ for $i = 1, 2, .., T$. Here n_i is the number of instances in class ω_i.

(ii) Find the sets of k nearest neighbors of Q from each class ω_i. In this case, the Minkowski distances are calculated from the training instances in $\{X_j, \omega_i\}_{j=1}^{n_i}$ to Q and the set of k nearest neighbors are identified from the reordered training instances according to the increasing distances.

(iii) For each set of k nearest neighbors $\{X_j^{nn}\}_{j=1}^{k}$ from each class ω_i (nn means nearest neighbor), power mean vectors M_i ($i = 1, 2, .., T$) are measured and which are called multi-local power mean vectors.

$$M_i = \left(\frac{1}{k}\sum_{j=1}^{k}(X_j^{nn})^p\right)^{1/p} \text{ for } p \neq 0 \tag{4}$$

(iv) Compute the Minkowski distances from Q to $M_i = \{\tilde{m}_1^i, \ldots, \tilde{m}_m^i\}$ for $i = 1, 2, .., T$ such as:

$$d_{Md}(Q, M_i) = \Big(\sum_{l=1}^{m} |q^l - \tilde{m}_i^l|^q\Big)^{1/q} \tag{5}$$

(v) Compute the memberships to $\{\omega_i\}_{i=1}^{T}$ according to Eq. (1) using the distances from Step (iv) and the crisp approach for calculating u_{ij} (i.e., $u_{ij} = 1$ for the known class and $u_{ij} = 0$ for other classes).

Algorithm 1 Updated MLPM-FKNN classifier

Input: $\{X_j, c_j\}_{j=1}^{n}$, k, p, q, Q
Output: ω^*
START
for $i \leftarrow 1$ to T **do**
  for $j \leftarrow 1$ to n_i **do**
    Compute $d_{Md}(Q, X_j) \leftarrow \big(\sum_{l=1}^{m} |q^l - x_j^l|^q\big)^{1/q}$
  end for
  Sort $\{d_{Md}(Q, X_j)\}_{j=1}^{n_j}$ in ascending order
  if ($n_j < k$) **then**
    $k \leftarrow n_j$
  end if
  Find $\{X_j^{nn}\}_{j=1}^{k}$
  Find $M_i \leftarrow (\frac{1}{k}\sum_{j=1}^{k}(X_j^{nn})^p)^{1/p}$
end for
for $i \leftarrow 1$ to T **do**
  Compute $d_{Md}(Q, M_i) \leftarrow \big(\sum_{l=1}^{m} |q^l - \tilde{m}_i^l|^q\big)^{1/q}$
  Compute $u_i(Q) \leftarrow \frac{\sum_{j=1}^{T} u_{ij}(1/d_{Md}(Q,M_i)^{2/(r-1)})}{\sum_{j=1}^{T}(1/d_{Md}(Q,M_i)^{2/(r-1)})}$
end for
return ω^* such that

$$\omega^* = \arg\max_{\omega_i} u_i(Q)$$

(vi) Classify Q into the class ω^* that has the highest membership degree. In other words:

$$\omega^* = \arg\max_{\omega_i} u_i(Q) \tag{6}$$

This method generates class-representative power mean vectors using k nearest neighbors obtained from each class subset instead of the entire training dataset. This distinguishes the proposed method from the original MLPM-FKNN algorithm. Moreover, utilizing the Minkowski distance metric to measure the distances from the query sample to the training instances allows the classifier to choose the most suitable distance metric based on the properties of the data. In the developed framework, we also examine the performance of the updated MLPM-FKNN classifier based on the Euclidean distance, which is denoted as MLPM-FKNN (E). At the same time, the Minkowski distance-based generalized approach is shown as MLPM-FKNN (M).

III. Data and Experimental Setting

A. Data Description

In this study, we used three datasets of biomass feedstocks, two of them were generated from several articles [7], [13], [14], [15], [16] published in respective journals, and other one was collected from the Phyllis 2 biomass data repository [17]. Information and the properties of each of the datasets are summarized in Table I. It is noteworthy to mention that these datasets are based on experimental outcomes of the proximate and ultimate analyses of biomass produced by previous studies. We attempt to use them for classification purposes in this study.

TABLE I: Properties of the data used

Data	Source	# Instances	# Features	# Classes
Dataset 1	[13], [14]	212	4	5
Dataset 2	[7], [15], [16]	135	5	5
Dataset 3	[17]	344	9	5

In these datasets, we included five classes of biomass feedstocks considering the property-based definitions in [5]. In particular, class 1 contained energy grasses and their parts (fiber materials, leaves), and class 2 comprised fruit residues and relevant sources (shells, seeds, pit). For class 3, materials from wood, wood chips, chips-barks, pruning were considered, while food crop residues (straws, stalks, dust, husk, hull, cob) were set for class 4. Class 5 included other waste materials such as milling industry waste, refuse, and municipal solid waste. Fig. 1 illustrates the percentages of the classes included in each dataset.

According to Fig. 1, it is clear that there are imbalances of the classes in each of datasets. Among them, class 3 is the most frequent class in all datasets, even though it does not account for over 50% of each dataset. In contrast, class 4 and 5 in dataset 1, and class 2 and 5 in dataset 2, and class 5 in dataset 3 are associated with a small number of biomass samples. The features considered in dataset 1 were fixed carbon (FC), volatile matter (VM), ash, and higher heating value (HHV) that had been extracted from the proximate analysis. In dataset 2, the features were the chemical properties of the biomass substances from the ultimate analysis, such as carbon (C), hydrogen (H), nitrogen (N), oxygen (O), and sulfur (S). For dataset 3, all feature types included in both dataset 1 and dataset 2 were considered. In all cases, we assumed that these features have significant influences on the class variable. Notice that dataset 2 and dataset 3 have not been used earlier for classification purposes, and this paper is the first one showing classification results for them. In particular, we utilize biomass data instances for the Phyllis 2 database for machine learning-based classification.

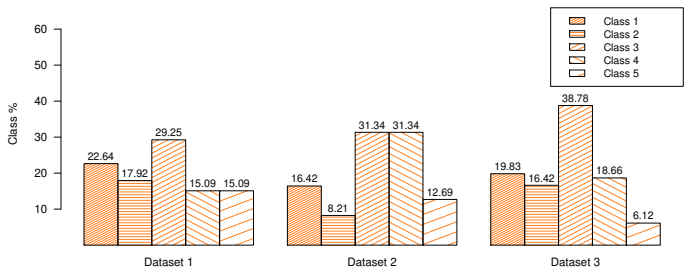


Fig. 1: Distribution (%) of each class in dataset 1, dataset 2, and dataset 3.

B. Testing methodology

The proposed framework for biomass classification has two main phases: i) training and validation and ii) testing. In the training and validation step, the model was developed by optimizing values for parameters k (number of nearest neighbors), p (power mean parameter), and q (Minkowski distance parameter). A grid search technique was deployed to optimize the model parameters. The performance of the classification models with optimal parameters were evaluated in the testing phase. To compare the performance of the generalized MLPM-FKNN classifier, we applied three well-known methods, namely k-nearest neighbor (KNN) [11], fuzzy k-nearest neighbor (FKNN) [18] and support vector machine (SVM) [19] classifiers. In addition to them, the MLPM-FKNN classifier based on the Euclidean distance [i.e., MLPM-FKNN (E)] was also applied, and the results were compared.

The analysis started with normalizing all features in the data into the unit interval. Next, datasets were randomly split into 60% for training, 20% for validation and 20% for testing. Stratified random sampling method was applied to ensure that all instances have the same proportions of units representing the different classes present as the whole data set. The holdout technique [26] was adopted for cross-validation, where the training and validation datasets were randomly generated 20 times. In the parameter settings, the number of nearest neighbors k was selected from $\{1, 2, ..., 15\}$ for all nearest neighbor methods. The value for p in power mean was chosen from the range $\{1, 1.1, .., 5\}$. The values from $\{1, 1.5, ..., 5\}$ were selected for the parameter q of the Minkowski distance. The fuzzy strength parameter $r = 2$ was kept, as in [12], [25] for MLPM-FKNN (M), MLPM-FKNN (E), and FKNN classifiers. Radial basis function kernel was used with the SVM model. To measure the classification performance, accuracy was used as the primary evaluation metric. Additional performance measures such as sensitivity and specificity were also measured as displaying classification results with accuracy alone is often not enough to adequately emphasize the effectiveness of the applied method [12]. The formulas used for sensitivity and specificity, especially to multi-class problems can be found from [25]. Additionally, the standard deviation (STD) of the accuracies was also computed. Based on the resulting confusion matrixes, we further examined the results of each classifier in the classification of biomass samples into each class.

IV. Results and Discussion

This section first presents the results from the training & validation phase of our methodology. Then the classification results in the test phase are presented.

A. Classification results with the training and validation data

We collected the accuracy, sensitivity, and specificity values in each run during the training and validation and averaged them for all repetitions from the holdout process. When the mean accuracy reached the maximum, the optimal values for the parameters (p, q and k) were observed. Table II

TABLE II: Classification performance with the validation data

Model	Measure	Dataset 1	Dataset 2	Dataset 3
MLPM-FKNN (Minkowski)	Accuracy	**0.5000**	**0.6217**	**0.7815**
	Sensitivity	0.4775	0.5208	0.7435
	Specificity	0.8697	0.8973	0.9447
	STD	0.0707	0.0761	0.0722
	Op. k, p, q	{9, 1.7, 1}	{2, 5, 3}	{3, 1, 1.5}
MLPM-FKNN (Euclidean)	Accuracy	0.4824	0.6152	0.7667
	Sensitivity	0.4558	0.5252	0.7175
	Specificity	0.8619	0.8968	0.9410
	STD	0.0676	0.0737	0.0636
	Op. k, p	{15, 2}	{2, 4.1}	{3, 1.4}
KNN	Accuracy	0.4588	0.5804	0.7370
	Sensitivity	0.4402	0.5402	0.6557
	Specificity	0.8582	0.8892	0.9317
	STD	0.0736	0.1183	0.0546
	Op. k	7	3	5
FKNN	Accuracy	0.4471	0.5804	0.7704
	Sensitivity	0.4313	0.5173	0.6839
	Specificity	0.8550	0.8866	0.9398
	STD	0.0676	0.0928	0.0500
	Op. k	15	11	6
SVM	Accuracy	0.4029	0.5348	0.7704
	Sensitivity	0.3600	0.3848	0.7056
	Specificity	0.8413	0.8684	0.9423
	STD	0.0312	0.0632	0.0211

summarizes those maximum performance measures and corresponding parameter values ("Op.") obtained with the proposed approach and the benchmarks with each dataset. To assess the reliability of the achieved mean accuracy value, its standard deviation ("STD") is also reported.

According to Table II results, we can see that the MLPM-FKNN (M) classifier achieves better results than the benchmarks in the training & validation for all datasets. It also has a reasonable standard deviation of accuracy and explicit support from mean sensitivity and specificity values. Moreover, used classifiers give outstanding performance with dataset 3 among all datasets while the proposed approach performs the best, achieving an accuracy of 78.15%. It is also apparent that the mean accuracy of all classifiers with dataset 2 is comparatively high compared with dataset 1, even though the sample size of dataset 2 is relatively small. This implies that the chemical properties of the biomass from the ultimate analysis offer great support than the proximate properties for their classifications, and having features from both analyses may provide even better results. Moreover, despite the influence of the class imbalance (as shown in Fig. 1) and the class overlapping issues [27], having a small number of instances in dataset 1, and dataset 2 might also have caused all classifiers to yield a relatively low performance.

Looking at the optimal values of the model parameters, a low value of k has yielded better results for MLPM-FKNN (M) than for the KNN and FKNN methods, which is surprising. This indicates that when the class-representative power mean vectors are computed using the k nearest neighbor from each class, it does not necessarily need to have more instances to make local power mean vectors more robust (and representative). It also can be seen that $p \in \{1.7, 5, 1\}$ and $q \in \{1, 3, 5\}$ have produced the maximum accuracy with the proposed MLPM-FKNN (M) approach for all datasets. Turning into the distance measure in the MLPM-FKNN classifier, the Minkowski distance-based approach has achieved slightly better accuracy than the Euclidean distance-based approach in all cases considered, which signifies the effectiveness of using Minkowski distance in the proposed method for biomass feedstock classification.

To visually inspect the impact of the different values of k and p on the classification performance of the proposed MLPM-FKNN (M) approach, Fig. 2 illustrates the mean accuracies during the training and validation with all datasets when q at its optimum.

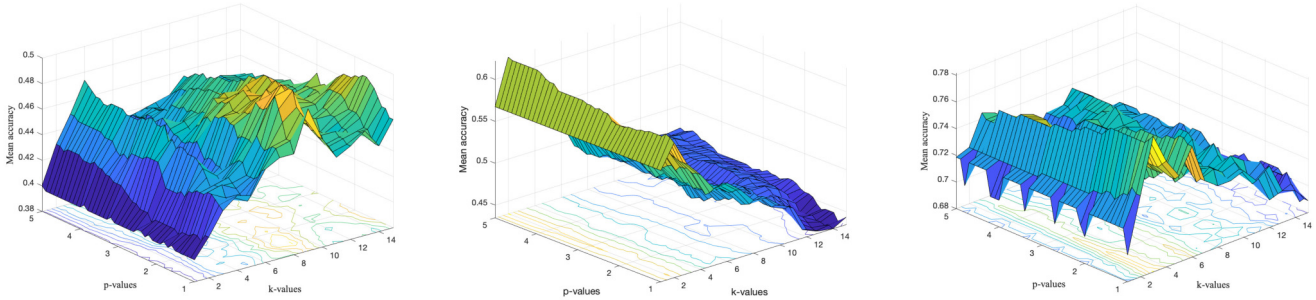
B. Classification performance with the test data

The classification results of each classifier with the test data instances are presented in Table III. In the testing step, we evaluated the performance of the trained models with the test data instances using the training instances that were stored during the holdout validation. As a result, the mean values of the performance measures are reported.

The results with the test data instances show that the proposed MLPM-FKNN (M) approach has a high classification accuracy compared to the benchmarks. In particular, it has a good accuracy of 70.88% with dataset 3, acceptable performance with dataset 2, and somewhat low accuracy of 42.62% with dataset 1. Along with them, other performance measures also remain reasonable, while the specificity is always higher

TABLE III: Classification performance with the test data

Model	Measure	Dataset 1	Dataset 2	Dataset 3
MLPM-FKNN (Minkowski)	Accuracy	**0.4262**	**0.5320**	**0.7088**
	Sensitivity	0.3850	0.4788	0.6935
	Specificity	0.8490	0.8736	0.9248
	STD	0.0508	0.0552	0.0161
MLPM-FKNN (Euclidean)	Accuracy	0.3786	0.5180	0.7059
	Sensitivity	0.3305	0.4721	0.6932
	Specificity	0.8349	0.8702	0.9242
	STD	0.0208	0.0527	0.0180
KNN	Accuracy	0.4000	0.4780	0.5853
	Sensitivity	0.3463	0.4538	0.5719
	Specificity	0.8442	0.8634	0.8931
	STD	0.0369	0.0458	0.0192
FKNN	Accuracy	0.3952	0.4760	0.6265
	Sensitivity	0.3508	0.3942	0.6232
	Specificity	0.8396	0.8570	0.9035
	STD	0.0256	0.0428	0.0305
SVM	Accuracy	0.4143	0.4960	0.6912
	Sensitivity	0.3605	0.3871	0.6721
	Specificity	0.8453	0.8591	0.9200
	STD	0.0392	0.0398	0.0180

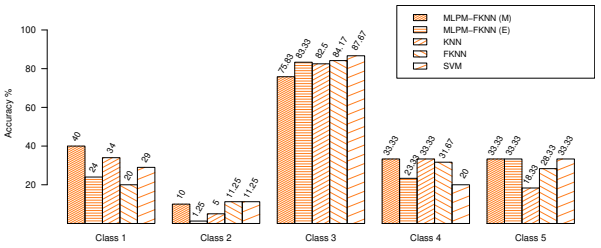


(a) Dataset 1 ($q = 1$) (b) Dataset 2 ($q = 3$) (c) Dataset 3 ($q = 1.5$)

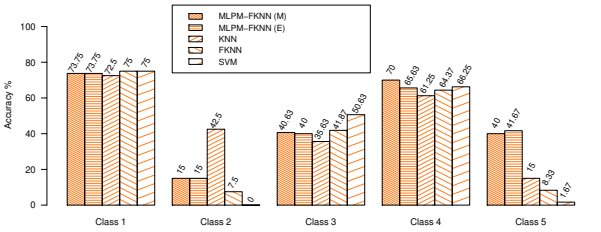
Fig. 2: Classification performance of the MLPM-FKNN (M) model with different p and k values for dataset 1, dataset 2, and dataset 3.

than the sensitivity. By looking at the others, even though the test performance of the KNN, FKNN and SVM models have comparable and generally good performance with dataset 3, they have relatively low performance with dataset 1 and dataset 2. Furthermore, it is apparent that for all methods used, the SDT is considerably lower for the test data (especially data set 3) than for the training and validation data.

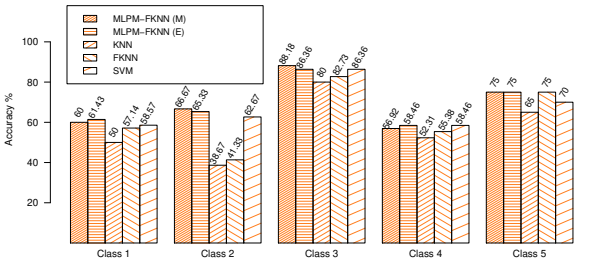
Fig. 3 shows the mean classification accuracy (measured from the confusion matrices) of each model for each class during the testing. It is apparent from the figure that all classifiers yielded good classifications on class 3 (that includes the wood-based energy crops) in dataset 1, whereas the SVM model performed the best. In dataset 2, class 1 (that includes energy grasses and their parts) and class 4 (that includes food crop residues-based biomass samples) have offered good and reasonable performance with all classifiers. In contrast, the classification performance of all methods in other classes of dataset 1 and dataset 2 appear to be poor—it is even worst for some cases, for instance, with class 2 in dataset 1. This might be because these classes are represented by a small number of biomass samples in the data. On the contrary, the classes (for example, class 3 in dataset 1) that are largely represented in the data have offered better classification. This indicates that the classification performance of these classes can be improved by introducing more data with approximately the same number of instances from all classes. It is also apparently supported by the results on dataset 3, where one can observe that the biomass samples in all classes generally produced good classification performance with all methods. This finding indicates that more biomass samples with relevant features from the proximate and ultimate analyses contribute to better results in their classification. Overall, it is evident from the result on dataset 3 that even though all the classifiers have comparable good performance, the MLPM-FKNN classifiers appear to be performing well for all classes classifications, whereas the KNN method performs the least.



(a) Dataset 1



(b) Dataset 2



(c) Dataset 3

Fig. 3: Comparison of classification performance of each model for each class with test data.

V. Conclusion

This paper presents a novel approach based on the MLPM-FKNN classifier and Minkowski distance for biomass feed-

stocks classification. An essential characteristic of this approach is that the generalization through power means and the Minkowski distance allows testing of different parameter values and enables a better fit of the method, consequently improving classification accuracy. We interpreted the biomass feedstocks classification as a five-class problem. Input features of the biomass samples included their characteristics from the proximate analysis and ultimate analysis. The experimental classification results clearly show that the proposed approach can achieve better performance than the benchmarks and can potentially produce an efficient classification that can benefit categorization of biomass sources for generating energy. The experimental results also validate the usefulness of the proposed MLPM-FKNN (M) method for multi-class imbalance real-world problems. Besides, it is evident from the results that the features from both ultimate and proximate analyses can offer a better classification of biomass feedstocks than the features considered from each of those analyses separately.

Future research possibilities include, for example, testing the classification performance of the proposed approach with more extensive biomass data that adequately comprises all classes specified in this study. Additional data will enhance the accuracy and the classification performance for wider range of biomass types and characteristics, in general.

REFERENCES

[1] A. Demirbas, "Biomass feedstocks," *In: Biofuels; Green Energy and Technology,* Springer, 2009, pp. 45-85.

[2] S. Gent and M. Twedt and C. Gerometta and E. Almberg, "Chapter Two - Introduction to Feedstocks," in *Theoretical and Applied Aspects of Biomass Torrefaction,* Butterworth-Heinemann, 2017, pp. 17-39

[3] A. A. Adeleke and J. K. Odusote and P. P. Ikubanni and O. A. Lasode, and M. Malathi, and D. Paswan,"The ignitability, fuel ratio and ash fusion temperatures of torrefied woody biomass," *Heliyon,* vol. 6, 2020, pp. e03582.

[4] A.A. Adeleke and P.P. Ikubanni and T.A. Orhadahwe and C.T. Christopher and J.M. Akano and O.O. Agboola and S.O. Adegoke and A.O. Balogun and R.A. Ibikunle, "Sustainability of multifaceted usage of biomass: A review," *Heliyon,* vol. 7, 2021, pp. e08025.

[5] O. O. Olatunji and S. Akinlabi and N. Madushele, "Property-based biomass feedstock grading using k-nearest neighbor technique," *Energy,* vol. 190, 2020, pp. 116346.

[6] P. Basu, Chapter 2 - Biomass Characteristics. *Biomass Gasification and Pyrolysis,* 2010, pp. 27-63.

[7] A. A. Khan and W. D. Jong and P. J. Jansens and H. Spliethoff, "Biomass combustion in fluidized bed boilers: Potential problems and remedies," *Fuel Process,* vol. 90, 2009, pp. 21-50.

[8] A. Nag and A. Gerritsen and C. Doeppke and A. E. Harman-Ware, "Machine Learning-Based Classification of Lignocellulosic Biomass from Pyrolysis-Molecular Beam Mass Spectrometry Data," *Int. J. Mol. Sci.,* vol. 22, 2021, pp. 4107.

[9] G. Tao and T. A. Lestander and P. Geladi and S. Xiong, "Biomass properties in association with plant species and assortments I: a synthesis based on literature data of energy properties," *Renew. Sustain. Energy Rev.,* vol. 16, 2012, pp. 3481-3506.

[10] M. Wang et al., "To distinguish the primary characteristics of agro-waste biomass by the principal component analysis: An investigation in East China," *Waste Manage.,* vol. 90, 2019, pp. 100-120.

[11] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory,* vol. 13, 1967, pp. 21-27.

[12] M. M. Kumbure and P. Luukka and M. Collan, "An enhancement of fuzzy k-nearest neighbor classifier using multi-local power means," *Proc. 11th Conf. European Society for Fuzzy Logic and Technology (EUSFLAT),* Atlantis Press. 2019, pp. 83-90.

[13] J. Parikh and S. A. Channiwala and G. K. Ghosal, "A correlation for calculating HHV from proximate analysis of solid fuels," *Fuel,* vol. 84, 2005, pp. 487-494.

[14] D. R. Nhuchhen and P. A. Salam, "Estimation of higher heating value of biomass from proximate analysis: A new approach," *Fuel,* vol. 99, 2012, pp. 55-63.

[15] S. V. Vassilev and D. Baxter and L. K. Andersen and C. G. Vassileva, "An overview of the chemical composition of biomass," *Fuel* vol. 89, 2010, pp. 913-933.

[16] M. Sajdak and O. Piotrowski, "C&RT model application in classification of biomass for energy production and environmental protection," *Cent. Eur. J. Chem.,* vol. 11, 2013, pp. 259-270

[17] Energy Research Centre of the Netherlands. *Phyllis 2: database for biomass and waste,* [Online]. Available: https://phyllis.nl/Browse/Standard/ECN-Phyllis#eucalyptus. [Accessed: July 31, 2021].

[18] J. M. Keller and M. R. Gray and J. A. Givens, "A Fuzzy K-Nearest Neighbor Algorithm," *EEE Trans. Syst. Man Cybern. Syst. ,* vol. 15, 1985, pp. 580-585.

[19] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.,* vol. 20, 1995, pp. 273-297.

[20] B. Salami and K. Haataja and P. Toivanen, "State-of-the-Art Techniques in Artificial Intelligence for Continual Learning: A Review" *Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems,* M. Ganzha, and L. Maciaszek, M. Paprzycki and D. Ślęzak, Eds. ACSIS, vol. 26, 2021, pp. 23-32.

[21] P. Gepner, "Machine Learning and High-Performance Computing Hybrid Systems, a New Way of Performance Acceleration in Engineering and Scientific Applications" *Proceedings of the 16th Conference on Computer Science and Intelligence Systems,* M. Ganzha, and L. Maciaszek, M. Paprzycki and D. Ślęzak, Eds. ACSIS, vol. 212, 2021, pp. 27-36.

[22] A. Coluccia and A. Fascista and G. Ricci, "A k-nearest neighbors approach to the design of radar detectors," *Signal Process.,* vol. 174, 2020, pp. 107609.

[23] R. Arian and A. Hariri and A. Mehridehnavi and A. Fassihi and F. Ghasemi, "Protein kinase inhibitors' classification using K-Nearest neighbor algorithm," *Comput. Biol. Chem.,* vol. 86, 2020, pp. 107269.

[24] S. Wua et al., "Evolving fuzzy k-nearest neighbors using an enhanced sine cosine algorithm: Case study of lupus nephritis," *Comput. Biol. Med.,* vol. 135, 2021, pp. 104582.

[25] M. M. Kumbure and P. Luukka and M. Collan, "A new fuzzy k-nearest neighbor classifier based on the Bonferroni mean," *Pattern Recognit. Lett.,* vol. 140, 2020, pp. 172-178.

[26] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Stat Surv.,* vol. 4, 2010, pp. 40–79.

[27] P. Vuttipittayamongkol and E. Elyan and A. Petrovski, "On the class overlap problem in imbalanced data classification," *Knowl.-Based Syst.,* vol. 212, 2021, pp. 106631