Article

An Ensemble Machine Learning Technique for Detection of Abnormalities in Knee Movement Sustainability

Hunish Bansal 1, Basavraj Chinagundi 1, Prashant Singh Rana 1* and Neeraj Kumar 1,2,*

1 Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala 147004, India
2 School of Computer Science, University of Petroleum and Energy Studies, Dehradun 248007, India
* Correspondence: neeraj.kumar@thapar.edu

Abstract: The purpose of this study was to determine electromyographically if there are significant differences in the movement associated with the knee muscle, gait, leg extension from a sitting position and flexion of the leg upwards for regular and abnormal sEMG data. Surface electromyography (sEMG) data were obtained from the lower limbs of 22 people during three different exercises: sitting, standing, and walking (11 with and 11 without knee abnormality). Participants with a knee deformity took longer to finish the task than the healthy subjects. The sEMG signal duration of patients with abnormalities was longer than that of healthy patients, resulting in an imbalance in the obtained sEMG signal data. As a result of the data’s bias towards the majority class, developing a classification model for automated analysis of such sEMG signals is arduous. The sEMG collected data were denoised and filtered, followed by the extraction of time-domain characteristics. Machine learning methods were then used for predicting the three distinct movements (sitting, standing, and walking) associated with electrical impulses for normal and abnormal sets. Different anomaly detection techniques were also used for detecting occurrences in the sEMG signals that differed considerably from the majority of data and were hence used for enhancing the performance of our model. The iforest anomaly detection technique presented in this work can achieve 98.5% accuracy on the light gradient boosting machine algorithm, surpassing the previous results which claimed a maximum accuracy of 92.5% and 91%, improving accuracy by 6–7% for classification of knee abnormality using machine learning.

Keywords: sEMG signal processing; machine learning; anomaly detection; feature selection

1. Introduction

Walking, running, or climbing stairs might be challenging if you have discomfort due to joint pain. Pain can arise as a result of a trauma or an injury. It may also start for no apparent cause at times. One study found that one in three people aged between 18 and 64 years have arthritis or joint symptoms due to an underlying injury or condition, such as osteoarthritis of the knee [1]. X-ray [2] or magnetic resonance imaging (MRI) methods [3] are used in the clinic to diagnose such abnormalities. The X-ray approach collects and assesses bone condition, whereas MRI gives comprehensive information about knee anatomies, such as cartilage, ligaments, and tendons. MRI is an effective diagnostic tool, but it is also more expensive. Wearable sensors, such as an electromyography (EMG), accelerometers, and gyro meters, and visual sensors such as imaging cameras can also be used to identify knee problems.

Depending on the parts of the human body supported by an exoskeleton, they can be divided into upper limb exoskeleton, lower limb exoskeleton (lees), whole body exoskeleton, and specific joint exoskeleton [4,5]. These exoskeleton systems are divided into three categories based on their various target applications and users, which are gait rehabilitation, human movement, and enhancement of human force [6].

Computer vision has applications across all areas of engineering, including exoskeletons. The exoskeleton obtains information about its environment from an RGB-D camera...
and extracts characteristics of the soil surface that could affect its gait. It then makes decisions according to the features of the environment, the state of the environment, and safety restrictions. It also provides the correct length and height of the walk to the parameterized gait planning model to assist the user in walking [7].

Sun et al. present the generation of exoskeleton mobility; several signal acquisition types can be applied, each focusing on specific areas and movements of the human body [8]. The current signal acquisition technologies include computer vision, electroencephalography (EEG), electromyography (EMG), and inertial sensors. These methods have been developed over the past several years. In addition, the detection and control method, stability of the modeling, and comfort of an exoskeleton of the lower limb will also impact its performance [8].

Calle et al. discussed the integration of inertial sensors into a lower limb robotic exoskeleton [9]. They present a novel platform based on Imocap-GIS inertial sensors with a motion-assistance exoskeleton. It generates joint movements via Maxon motors and harmonic drive gearboxes.

Tang et al. present an extensive study on a wearable lower-limb exoskeleton [10]. They classified the wearable device according to the power source and the working principle. They also compared and analyzed the design idea, wearing mode, material, and performance of different types of lower-limb exoskeletons.

Taylor et al. demonstrate real-time muscle length tracking in an in vivo turkey model via chronically implanted magnetic beads while investigating accuracy, biocompatibility, and long-term implant stability [11].

Botter et al. combined ultrasound (US) imaging and multichannel surface EMG to compare their detection sensitivity to fasciculations occurring in different muscle regions and to investigate the effect of EMG electrodes’ configuration on their agreement [12]. Monopolar surface EMGs were collected from medial gastrocnemius and soleus with an array of 32 electrodes.

EMG sensors recognize motions in advance while detecting signal changes faster, making them better at researching neuromusculoskeletal diseases. The signal obtained by the EMG sensors is a biological signal which measures the electrical activity produced by the skeletal muscles. Surface electromyogram (sEMG) and intramuscular EMG (iEMG) techniques are important methods to acquire EMG signals [13]. sEMG has an advantage over iEMG in that the electrodes may be placed without causing discomfort or requiring medical supervision, and the risk of infection is minimum.

Gui et al. proposed an adaptive estimator based on the EMG to obtain and update the EMG pair model without calibrations and recalibrations [14]. Simulation and experiments indicate that the proposed estimator can adaptively predict the subject’s active joint couple and provide precise control of exoskeleton movement.

Kang et al. developed a neural network gait and slope estimator using electromyogram (EMG) and mechanical sensor signals [15]. The results of four healthy and two elderly subjects show that the EMG approach can reduce the error rate by 14.8% compared to models using only mechanical sensors.

Long-term monitoring with surface electrodes is simpler than with iEMG needles. As Kanoga et al. have shown, the location of the sEMG sensors has a considerable impact on the collected signal and its subsequent analysis and identification using the calculation method [16]. sEMG sensors can collect data during daily human activities, such as sitting, climbing, walking, and standing, which can then be used to recognize actions, detect anomalies, etc. The goniometer was connected out of the knee joint, and sEMG data were collected for four muscles: rectus femoris (RF), biceps femoris (BF), vastus medialis (VM), semitendinosus (ST), and semimembranosus (SM) (Figure 1).
Knee pain has become one of the most common health issues among seniors. Wounds, repeated stress on the joint, aging, and underlying conditions such as arthritis are the most common causes of knee pain [17]. The knee joint is one of the most complex joints in the human body, ensuring leg movement and body stability and acting as a damper [18]. The knee joint is made up of different bones, tendons, muscles, cartilage, ligaments, and fluids. During the past decade, researchers have focused on categorizing sEMG signals from the upper limb. The sEMGs obtained from the lower limb are more difficult to interpret because they are influenced by overlapping muscles [17,18]. Various signal processing approaches [19,20] have been used to study the categorization of sEMG data from the lower leg. Chen et al. used a deep-neural-network to predict lower-limb extension/flexion joint angles from sEMG signals [21]. Morbidoni et al. used an in-depth learning method to classify the walking step using sEMG data [22]. Bonato et al. examined the depletion of quadriceps and hamstring muscles using sEMG signals [23]. The classification of walking patterns and the use of computational approaches are divided into normal and gait disorders [24].

1.2. Contributions

This paper focuses on identifying abnormality in movements from imbalanced sEMG data and classifying six different activities (three normal and three abnormal) for standing, sitting, and gait positions. We thus present a performance comparison of various prediction classifiers for detecting knee abnormalities from sEMG signals in which classifiers are applied on the datasets consisting of data with and without anomalies in using multiple anomaly detection techniques. This paper aims to enhance the ability to detect knee movement abnormalities and enable healthcare for a more accurate diagnosis. To the author’s knowledge, no study solves the problem of identifying knee abnormalities using anomalous detection methods. The major contributions are described as follows:

- Anomaly detection techniques such as iforest, KNN, and lof have been used to remove abnormal data.
- Enhancement of classifier’s accuracy using boosting algorithms.

1.3. Organization

In this paper, Section 2 describes our proposed model’s datasets, methodology, and workflow. Section 3 discusses the model evaluation parameters. Section 4 defines the performance and provides a comparative analysis between different metrics. Finally, the paper is concluded in Section 5.

2. Materials and Methods

2.1. Dataset and Its Features

The sEMG and joint angles were derived from the publicly available dataset at UCI machine learning repository [25]. The data includes sEMG signals from the lower limbs of 22 adults, 11 of whom are in good health and 11 of whom have known knee problems.
Participants were in good health with no history of injury or knee discomfort. One abnormal subject of the knee suffered a sciatic nerve lesion, six subjects suffered an anterior cruciate ligament (ACL) lesion, and the other four suffered a meniscus lesion. Dataset description and feature information have been provided in Tables 1 and 2, respectively. These data were collected using a Biometrics Ltd. DataLog MWX8 and a direction finder. Each person performed three distinct tasks: walking, bending the leg upward, and lengthening the leg from a sitting posture. The goniometer was connected out of the knee joint, and sEMG data was collected for four muscles: vastus medialis (VM), rectus femoris (RF), semitendinosus (ST), and biceps femoris (BF) (Figure 1).

Table 1. Description of the dataset.

| Description                          | Type/Count |
|--------------------------------------|------------|
| Number of features                   | 5          |
| Feature characteristics              | Real       |
| No. of classes (Movements)           | 6          |
| Dataset characteristics              | Multivariate |
| Associated tasks                     | Classification |
| Missing values                       | No         |

Table 2. Description of the features.

| Feature | Information       |
|---------|-------------------|
| RF      | Rectus Femoris    |
| BF      | Biceps Femoris    |
| VM      | Vastus Medialis   |
| ST      | Semitendinosus    |
| FX      | Flexion           |

The person’s afflicted limb with aberrant knee and left leg healthy participants were chosen to receive the signal. The data were collected at a sampling frequency of 1000 Hz with a resolution of 14 bits. The sEMG signals had already been filtered through a band-pass filter ranging from 20 Hz to 460 Hz. The data gathered do not have a signal corresponding to the transition steps, such as standing, sitting, walking, standing, etc. In Table 3, movements labeled as 0 refer to normal walking, 1 for normal sitting, 2 for normal standing, 3 for abnormal walking, 4 for abnormal sitting, and 5 for abnormal standing, respectively. The data were transmitted in real-time through Bluetooth from the MWX8 gadget to the PC.

Table 3. Sample dataset.

| RF      | BF       | VM       | ST       | FX       | Movement |
|---------|----------|----------|----------|----------|----------|
| 0.0011  | -0.0011  | 0.0021   | 0.0050   | 57.6     | 0        |
| 0.0011  | -0.0010  | 0.0022   | 0.0052   | 57.5     | 0        |
| 0.0011  | -0.0009  | 0.0024   | 0.0054   | 57.3     | 0        |
| 0.0051  | 0.0002   | -0.0010  | 0.0007   | 3        | 1        |
| 0.0052  | 0.0002   | -0.0007  | 0.0007   | 3        | 1        |
| 0.0056  | 0.0003   | -0.0001  | 0.0007   | 3        | 1        |
| 0.0011  | 0.0002   | 0.0002   | 0.0000   | 70.1     | 2        |
| 0.0011  | 0.0002   | 0.0003   | 0.0000   | 70.1     | 2        |
| 0.0011  | 0.0002   | 0.0004   | 0.0000   | 70.1     | 2        |
| -0.0004 | 0.0002   | 0.0184   | 0.0017   | -9       | 3        |
Table 3. Cont.

| RF   | BF   | VM   | ST   | FX   | Movement |
|------|------|------|------|------|----------|
| 0.0004 | 0.0011 | 0.0187 | 0.0022 | −9   | 3        |
| 0.0012 | 0.0021 | 0.0190 | 0.0028 | −9   | 3        |
| 0.0011 | 0.0018 | 0.0008 | 0.0012 | −4.6 | 4        |
| 0.0011 | 0.0019 | 0.0009 | 0.0013 | −4.5 | 4        |
| 0.0012 | 0.0020 | 0.0009 | 0.0013 | −4.6 | 4        |
| 0.0043 | 0.0013 | −0.0019 | 0.0005 | −30  | 5        |
| 0.0045 | 0.0013 | −0.0017 | 0.0005 | −29.7| 5        |
| 0.0048 | 0.0014 | −0.0015 | 0.0005 | −29.4| 5        |

Statistical Analysis of Dataset

Table 4 and Figure 2 show the statistical analysis of all the datasets. The standard deviation of FX is the highest, and BF is the lowest in all the datasets. EMG signal plots can be seen before data preprocessing in Figure 2a for the four channels of a normal person and in Figure 2b for an abnormal person. EMG signal plots can be seen after signal preprocessing, in which we rectify and transform the filtered sEMG values we can see in Figure 2c for a normal person and in Figure 2d for an abnormal person. After normalizing, we achieve the plots as shown in Figure 2e for a normal person and in Figure 2f for an abnormal person.

Table 4. Statistical analysis of each feature in each dataset.

| Dataset   | Features | Mean. | Std.  | Min. | Max.  |
|-----------|----------|-------|-------|------|-------|
| Nor.Walk  | RF       | 0.04  | 0.016 | −0.11| 0.10  |
|           | BF       | −0.02 | 0.008 | −0.06| 0.11  |
|           | VM       | −0.01 | 0.026 | −0.18| 0.17  |
|           | ST       | −0.03 | 0.011 | −0.08| 0.11  |
|           | FX       | 19.36 | 17.836| 1.30 | 62.60 |
| Abn.Walk  | RF       | −0.01 | 0.083 | −1.27| 1.36  |
|           | BF       | −0.06 | 0.023 | −0.32| 0.16  |
|           | VM       | −0.06 | 0.068 | −0.61| 0.59  |
|           | ST       | 0.06  | 0.016 | −0.15| 0.10  |
|           | FX       | −12.7 | 16.37 | −57.2| 3.10  |
| Nor.Sit   | RF       | 0.03  | 0.016 | −0.11| 0.10  |
|           | BF       | −0.02 | 0.008 | −0.06| 0.11  |
|           | VM       | −0.01 | 0.026 | −0.18| 0.17  |
|           | ST       | −0.03 | 0.011 | −0.08| 0.11  |
|           | FX       | 53.36 | 36.836| 3.30 | 109.6 |
| Abn.Sit   | RF       | −0.03 | 0.006 | −0.08| 0.05  |
|           | BF       | −0.06 | 0.034 | −0.33| 0.20  |
|           | VM       | −0.04 | 0.012 | −0.11| 0.20  |
|           | ST       | 0.05  | 0.06  | −0.56| 0.40  |
|           | FX       | −58.7 | 44.37 | −113 | −2.90 |
| Nor.Stand | RF       | 0.03  | 0.018 | −0.11| 0.10  |
|           | BF       | −0.03 | 0.003 | −0.06| 0.11  |
|           | VM       | −0.01 | 0.026 | −0.18| 0.17  |
|           | ST       | −0.04 | 0.011 | −0.08| 0.11  |
|           | FX       | 51.36 | 20.836| 20.30| 75.6  |
Table 4. Cont.

| Dataset    | Features | Mean  | Std.  | Min.   | Max.   |
|------------|----------|-------|-------|--------|--------|
|            | RF       | −0.01 | 0.032 | −0.20  | 0.26   |
|            | BF       | −0.06 | 0.006 | −0.05  | 0.06   |
|            | VM       | −0.06 | 0.019 | −0.16  | 0.14   |
|            | ST       | 0.07  | 0.02  | −0.49  | 0.35   |
|            | FX       | −29.4 | 30.37 | −83    | −4.70  |
| Abn.Stand  |          |       |       |        |        |
|            | RF       | −0.01 | 0.032 | −0.20  | 0.26   |
|            | BF       | −0.06 | 0.006 | −0.05  | 0.06   |
|            | VM       | −0.06 | 0.019 | −0.16  | 0.14   |
|            | ST       | 0.07  | 0.02  | −0.49  | 0.35   |
|            | FX       | −58.7 | 44.37 | −113   | −2.90  |

(a) Raw sEMG signal for normal knee
(b) Raw sEMG signal abnormal knee
(c) Hilbert processed and rectified sEMG signal for normal knee
(d) Hilbert processed and rectified sEMG signal abnormal knee
(e) Normalized sEMG signal for normal person
(f) Normalized sEMG signal abnormal knee

Figure 2. The sEMG signals at various stages during data preprocessing.

2.2. Approach

The approach is described in Figure 3. Firstly, the raw experimental data of the lower limb is taken from the UCI machine learning repository. The dataset consists of data for three different activities (sitting, standing, and walking) for 11 abnormal and 11 normal individuals labeled as Abn. And Nor. In Table 4.
2.2. Approach

The approach is described in Figure 3. Firstly, the raw experimental data of the lower limb is taken from the UCI machine learning repository. The dataset consists of data for three different activities (sitting, standing, and walking) for 11 abnormal and 11 normal individuals labeled as Abn. And Nor. in Table 4.

The data is pre-filtered with a band-pass filter that ranges from 25 Hz to 465 Hz. The signal data is later demodulated using Hilbert’s transformation. In case of mechanical failure, the gathered vibration signals are usually modified. As such, signal demodulation can separate the carrier component from the modulation component. In addition, the default features are still hidden in the modulation component. Thus, implementation of signal demodulation during signal preprocessing is essential for precise results. The min/max normalization method was applied to normalize the samples with features ranging from 0 to 1, as shown in Figure 3. The dataset is then divided into two groups:

1. Dataset consisting of anomalies
2. Datasets produced as a result of the use of anomaly detection techniques, namely iforest, KNN, and lof.

Following the generation of these datasets, we use machine learning classifiers that comprise of models:

- Light gradient boosting machine
- Extra trees classifier
- Random forest classifier
- Extreme gradient boosting
- Decision tree classifier

Figure 3. Methodology used.
2.2.1. Anomaly Detection Techniques

Anomaly detection techniques (also known as outlier analysis) is a stage in data mining that finds data points, events, and/or observations that deviate from the normal behavior. Anomalous analytics may indicate major problems, such as a technical flaw or potential scenarios. We analyze the sEMG signals and remove outliers from them using the following techniques:

1. iforest: The Isolation Forest is an unsupervised learning algorithm to detect abnormalities that operate on the principle of isolating abnormalities rather than the most common staining profiling techniques [26]. The decision tree algorithm backs the secluded forest. It separates aberrant values by randomly selecting a character from the given set of characteristics and then randomly selecting a value divided between the max and min values. This random distribution of traits will lead to shorter tree trajectories for anomalous data points, differentiating them from the rest of the data. Generally, the first step in detecting an anomaly is to profile what is “normal” and then flag anything that cannot be considered typical as abnormal. However, the forest isolation algorithm does not function on this principle; it first defines the “normal” behavior and calculates the distances per point. As expected, the isolation forest functions instead by isolating anomalies by explicitly isolating abnormal issues in the dataset. The isolation forest algorithm assumes that specific abnormalities and observations should make it easier to identify them. The isolated forest uses a series of trees to isolate anomalies for the given data points [26]. The isolation forest recursively generates scores across the dataset by randomly selecting a feature and then a split value. The monsters likely need fewer random partitions than the “normal” points of the dataset. The faults will be the dots with a smaller path into the tree and the crossed edges of the root node.

2. KNN (k-Nearest Neighbors Detector): The unsupervised k-nearest-neighbor global 201 anomaly detection technique is an easy way to detect anomalies and not be confused with the k-nearest classification [27]. As the name suggests, it focuses on global anomalies and cannot see local ones. First, the closest k-neighbors must be found for every record in the data set. Afterward, an anomaly score is calculated using those neighbors [27]. On the other hand, two possibilities have been suggested: either the distance to the nearest kth-neighbor is used (only one), or the average length to all the nearest k-neighbors is calculated. In the subsequent paragraphs, we refer to the first method as kth-NN and the second as k-NN. In practice, the k-NN method is usually preferred. However, the absolute value of the score depends a lot on the data set itself, the number of dimensions, and the standardization [27]. Therefore, it is not easy to choose an appropriate threshold, if necessary. Naturally, selecting the parameter k is essential to the results. If too low, the density estimate for the records could not be reliable. In addition, the density estimate may be too coarse if it is too large. Usually, k should be in the range $10 < k < 50$. In classification, a suitable k can be determined, for example, using cross-validation. Unfortunately, there is no such technique for detecting unsupervised abnormalities due to missing tags. For this reason, we later evaluate many different values for k and mean to obtain a fair assessment when comparing algorithms.

3. LOF (Local Outlier Factor): The LOF uses the identification of density-based outliers to identify local outliers or sites that are outliers relative to their immediate environment rather than the overall distribution of data [28]. The higher the LOF value of an observation, the more astonishing it becomes. The number of regarded neighbors is usually fixed to be: (1) less than the maximum number of samples close by which may be local outliers and (2) exceeds the minimum number of samples that a cluster must contain for other samples to be local outliers [28].
2.2.2. Feature Importance

Techniques that score each feature according to their usefulness in predicting a target variable are known as feature importance. Statistical correlation scores, coefficients derived from linear models, decision trees, and permutation significance scores are a few examples of trait importance scores. Feature significance scores are essential for predictive modeling projects. They provide an overview of the data and information on the template. A foundation for a reduction in dimensions and selection of characteristics can improve the effectiveness of a predictive model on the problem.

The importance of the feature is computed using Equation (1). After five different runs, the average weight given to each feature is described in Table 5 and their average weight is used for ranking the features. It is found that FX has the highest ranking and RF has the lowest ranking.

\[
    \text{Rank} = \sum_{i=1}^{C} -f_i \log(f_i) \quad (1)
\]

where, \( C \) is the unique labels in the dataset, \( f_i \) is the frequency if label \( i \) at a node.

Table 5. The importance of each feature.

| Runs | FX  | BF  | ST  | VM  | RF  |
|------|-----|-----|-----|-----|-----|
| 1    | 0.25| 0.22| 0.2 | 0.19| 0.14|
| 2    | 0.23| 0.2 | 0.24| 0.17| 0.16|
| 3    | 0.27| 0.23| 0.19| 0.16| 0.15|
| 4    | 0.25| 0.21| 0.21| 0.17| 0.16|
| 5    | 0.24| 0.2 | 0.2 | 0.17| 0.17|
| Avg. | 0.248| 0.212| 0.208| 0.172| 0.156|

2.2.3. Computational Classifiers

The dataset was divided into 75/25 Train-Test halves. The resulting datasets are subsequently trained on a variety of machine learning classifiers. The results are compared between the top five performing classifiers:

- Light gradient boosting machine
- Extra trees classifier
- Random forest classifier
- Extreme gradient boosting
- Decision tree classifier

1. LightGBM: This improves gradient amplification by incorporating automatic feature selection and focusing on amplification instances with higher gradients [29]. This will lead to much faster workout times and greater predictive efficiency. It is known to be a very effective computational algorithm and rapid treatment method. As the trees of other algorithms grow horizontally, LightGBM grows vertically, in the direction of the leaves, while different algorithms grow in level. LightGBM picks the sheet with the highest loss to develop. When culturing the same sheet, it will decrease the loss more than a leveling method. Decision trees are used to improve model performance while saving memory. Quality and scalability are at the heart of the development approach. It uses two new techniques: one-sided gradient sampling and proprietary feature grouping (EFB). Fill in the histogram algorithm gaps in most GBDT (gradient boosting decision tree) frameworks [29]. The LightGBM algorithm’s characteristics are produced by the gradient-based one side sampling technique (GOSS) and the exclusive feature bundling technique (EFB). They work together to make the model work and give it an edge over the competition for GBDT systems.

2. Extreme gradient boosting: Unlike traditional gradient boosting, XGBoost uses its tree-building technique where similarity score and gain determine optimal node divisions [30]. Residual value: real value (observed)-predicted value. The probability
of an event calculated in an earlier stage is called an earlier probability. For each finding, the initial probability is considered to be 0.5, which is used to create the first tree. The process map shows that the previous probability is reconstructed for all future trees based on the original predictions and the predictions of all previous trees. Lambda is a setting for equalization. The increase in lambda diminishes the influence of small leaves (those with few observations) disproportionately while having a moderate impact on large leaves (those with many observations) [30]. The optimum division for the tree is then determined by selecting the division knot with the greatest gain.

3. Random Forest: A random forest is a holistic approach that uses multiple decision trees [31]. It is also known as bootstrapping and aggregation, often called bagging, to perform regression and filing tasks. The simple hypothesis is that instead of relying on individual decision trees, the outcome is determined by combining multiple decision trees. Random Forest uses a variety of decision trees as a primary learning approach. The dataset is randomly sampled for line and feature sampling, resulting in sample data sets for each model. This section is called bootstrap. Random logging is a method of bagging and not a method of stimulation. Trees in random forests grow next to each other. Growing trees are not related. A randomized forest is a meta-marker (it incorporates the results of several predictive models) that groups together many decision trees with some helpful changes. It is one of the most precise machine learning technologies available on the market. On many datasets, this is a very reliable grader. It works effectively on big datasets [31]. It can handle tens of thousands of input variables without removing them. It identifies the importance of a large number of factors in categorization.

4. Extra Trees: This module presents a meta-estimator that uses the mean to improve statistical accuracy [32]. It combines a range of randomized decision trees. Extra trees is an intuitive learning approach that combines the predictions of multiple decision trees into a single prediction (the random forest model previously discussed). It may produce similar or better results than the random forest technique because it uses a simpler algorithm to create the entire decision trees. Following the standard upper-level approach, the extra-trees algorithm produces an arbitrary decision or a set of regression trees. It differs from earlier tree-based methods by randomly dividing nodes and generating trees using the whole learning sample (rather than a bootstrap replication) [32].

5. Decision Tree: Decision tree learning, also known as decision tree induction, is one of the statistical modeling methods used in analytics, data processing, and machine learning [33]. It uses a decision tree from assumptions about an item to predictions about the article’s target value. Classification trees represent class tags, and branches represent combinations of features that match specific class tags. A decision tree is a graphic and explicit depiction of the options and decision making that can be used in decision-making procedures. A decision tree is a data model for exploring the data. The decision tree learning approach is often used in data mining. The objective is to build a model that can predict the value of a target variable based on a set of input variables. The decision tree method is a supervised learning approach that can generate continuous and categoric production variables.

3. Model Evaluation

There are many classification metrics for evaluation of our model. In this study, we assess the performance on (1) Matthews correlation coefficient (MCC), (2) F1 Score (F1), (3) accuracy, (4) precision, and (5) AUC.

3.1. Matthews Correlation Coefficient (MCC)

The MCC formula takes into account all of the cells in the confusion matrix.

\[
\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]
Similar to the correlation coefficient, MCC has a range of values ranging from \(-1\) to +1. A model with a score of +1 is considered ideal, whereas a model with a score of \(-1\) is considered bad. This feature is one of the most important benefits of MCC since it leads to simple interpretation.

3.2. F1 Score

The F or F score measurement is a measurement of test accuracy. It is calculated with the accuracy and recalculation values from the test. Accuracy is the number of actual positive outcomes divided by the total number of positive outcomes, including those incorrectly identified. The recall is the number of true positive results divided by the total number of samples which should have been positive.

\[
F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3}
\]

3.3. Precision

Precision, sometimes referred to as the positive predictive value, is computed as,

\[
Precision = \frac{TP}{TP + FP} \tag{4}
\]

Precision may produce a biased result since the formula lacks FN and TN, especially for unbalanced classes.

3.4. Accuracy

Accuracy is a criterion for classifying models. Informally, precision is the percentage of accurate predictions from our model. Formal accuracy is defined by:

\[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}
\]

where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives.

3.5. AUC

AUC is an abbreviation for “Area under the ROC Curve.” AUC measures the whole two-dimensional area beneath the entire ROC curve from \((0,0)\) to \((1,1)\). AUC is a metric that aggregates performance across all categorization criteria. AUC may be interpreted as the likelihood that the model rates a random positive case higher than a random negative example.

4. Result Analysis

Here, we examine the prediction outcomes of five machine learning classifiers on the training, testing, and validation dataset. The performance of each model is shown in Table 6. All the prediction models have been trained on default parameters and evaluated on MCC, accuracy, F1, precision, and AUC. The dataset has a small number of features but a large number of observation values. The K-fold cross-validation technique is used to assess the robustness of the best prediction method.
### Table 6. Performance comparison of different anomaly detection techniques with different prediction models on different metrics on all the datasets.

| Anomaly Detection Technique | Models → | lgbm | xgboost | rf | dt | et |
|-----------------------------|----------|------|---------|----|----|----|
|                            | Metrics ↓ |      |         |    |    |    |
| Without Anomaly Detection   | MCC      | 0.83 | 0.753   | 0.559 | 0.537 | 0.367 |
|                             | AUC      | 0.972 | 0.958 | 0.897 | 0.831 | 0.839 |
|                             | Precision | 0.859 | 0.798 | 0.613 | 0.514 | 0.59 |
|                             | F1       | 0.857 | 0.79 | 0.617 | 0.522 | 0.442 |
|                             | Accuracy | 0.859 | 0.797 | 0.64 | 0.537 | 0.478 |
| iForest                     | MCC      | **0.985** | 0.982 | 0.922 | 0.822 | 0.816 |
|                             | AUC      | **0.986** | 0.986 | 0.983 | 0.967 | 0.974 |
|                             | Precision | **0.985** | 0.984 | 0.937 | 0.859 | 0.865 |
|                             | F1       | **0.985** | 0.983 | 0.933 | 0.855 | 0.843 |
|                             | Accuracy | **0.985** | 0.983 | 0.935 | 0.856 | 0.852 |
| KNN                         | MCC      | 0.922 | 0.9 | 0.785 | 0.569 | 0.72 |
|                             | AUC      | 0.983 | 0.98 | 0.948 | 0.862 | 0.918 |
|                             | Precision | 0.956 | 0.946 | 0.893 | 0.8 | 0.861 |
|                             | F1       | 0.954 | 0.942 | 0.884 | 0.765 | 0.854 |
|                             | Accuracy | 0.953 | 0.941 | 0.88 | 0.748 | 0.853 |
| lof                         | MCC      | 0.908 | 0.805 | 0.558 | 0.454 | 0.43 |
|                             | AUC      | 0.979 | 0.962 | 0.894 | 0.837 | 0.834 |
|                             | Precision | 0.926 | 0.847 | 0.692 | 0.6 | 0.631 |
|                             | F1       | 0.926 | 0.847 | 0.653 | 0.571 | 0.506 |
|                             | Accuracy | 0.926 | 0.847 | 0.666 | 0.583 | 0.569 |

### 4.1. Model Performance Comparison

After training models on different datasets, it is observed that the dataset generated using iForest performs the best overall on the light gradient boosting machine model with an impactful accuracy of 0.985. Other anomaly detection technique-generated datasets also perform better than the initial dataset, with an accuracy of 0.859. The best model is validated using 10-fold validation.

#### 4.1.1. Accuracy

The accuracy is calculated using Equation (5). Figure 4a shows the average accuracy of all the models. The light gradient boosting machine has the highest accuracy of 0.985, followed by xgboost and random forest. Extra trees have the lowest accuracy of 0.478.
4.1.2. F1 Score

The F1 is calculated using Equation (3). Figure 4b shows the average F1 score of all the models after 10-fold validation. The light gradient boosting machine has the highest F1 of 0.985, followed by xgboost and random forest. Extra trees have the lowest accuracy of 0.442.

4.1.3. MCC

The MCC is calculated using Equation (2). Figure 4c shows the average MCC of all the models after 10-fold validation. The light gradient boosting machine has the highest accuracy of 0.985, followed by xgboost and random forest. Extra trees have the lowest accuracy of 0.367.

4.2. K-Fold Validation

Cross-validation is a statistical method for estimating capacity in machine learning models. It is commonly used in applied machine learning to compare and select a model for a specific predictive modeling problem. It is easy to understand, implement, and produce skill estimates with a more negligible bias than other approaches. Cross-validation is a resampling technique used to assess ML models on a small sample of data. The process contains only one parameter called k, which specifies the number of groups in which a given data sample should be divided.

Consequently, this method is often referred to as k-fold cross-validation. When a specific number for k is specified, it may be used instead of k in the model reference, such as k = 10, resulting in a 10-cross-validation. Cross-validation is primarily used in applied machine learning to assess the competency of a machine learning model on unseen data. It uses a small sample to determine how the model will work when used to produce projections of data not used in model formation. The charts show that the built model is robust after 10 plies and is reliable for testing our sEMG data.
4.3. Results and Discussion

The results indicate that datasets created following the anomaly detection techniques increase the performance of classifiers for movement prediction.

The performance of each model is shown in Table 6. All the prediction models have been trained on default parameters and evaluated on MCC, accuracy, F1, precision, and AUC. The dataset has a small number of features but a large number of observation values. The K-fold cross-validation technique is used to assess the robustness of the best prediction method.

The dataset created with iforest achieves an outstanding overall score on the light gradient boosting machine model with an accuracy of 98.5% compared to the baseline performance without any adjustments on our dataset, which achieves an accuracy of 85.9%. The other models significantly improve their accuracy and the metrics after anomaly removal from the data. There is a performance increase of 13% accuracy from the models which do not use anomaly detection techniques. Compared to previous works [34,35] which were able to achieve an accuracy of 92.5% and 91%, this has also been improved by 6–7%, respectively, which proves that this technique is effective in boosting the accuracy for classification of sEMG signals. Thus, results demonstrate that eliminating anomalies substantially improves performance, producing some robust results, and should be used for developing models using sEMG data.

4.4. Limitation of the Study

Data imbalance is an essential issue in the classification of medical data. This can occur because of significant differences in the number of healthy and unhealthy individuals. This may happen because the data collection length varies depending on normal and abnormal subjects. An individual with an anomaly of the knee takes longer to accomplish the task of moving, resulting in a more extended signal. Therefore, the abnormal subject’s sEMG signal length results in a class imbalance. This affects the precision of the classification of knee motion and the prediction of anomalies.

5. Conclusions

This work demonstrated machine learning for movement classification using sEMG data. The suggested models are based on a large database with various parameters. This study aimed to analyze differences in classification performance following anomalous detection methods. It was used to identify events in EMG signals that differed significantly from most data and to create new data sets.

Real-time sEMG data analysis and movement classification can be beneficial in the medical field. Doctors could identify aberrant motions and provide preventative steps to minimize risks and injuries connected to joints or muscles. This work hopes to contribute towards a more accurate diagnosis system to predict abnormalities and suggest precautionary measures to prevent any unfortunate events. In future research, we plan to gather DGG data from other topics to reduce data imbalances, to test oversample methods, and study the most advanced machine learning graders. Convolution neural networks (CNN) have shown promising results in various classification questions. The iforest anomaly detection techniques presented here can be up to 98.5% accurate with light gradient boosting. It enhances the ability to detect defects within EMG signals. The proposed methodology can be used in muscular deficiency or any other physiological deficiency.

Author Contributions: Conceptualization, N.K.; Data curation, H.B. and B.C.; Formal analysis, B.C. and N.K.; Methodology, H.B., B.C., P.S.R. and N.K.; Validation, H.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.
29. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
30. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K. Xgboost: Extreme gradient boosting. R Package Version 2015, 1, 1–4.
31. Biau, G.; Scornet, E. A random forest guided tour. TEST 2016, 25, 197–227. [CrossRef]
32. Ahmad, M.W.; Reynolds, J.; Rezgui, Y. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. J. Clean. Prod. 2018, 203, 810–821. [CrossRef]
33. Myles, A.J.; Feudale, R.N.; Liu, Y.; Woody, N.A.; Brown, S.D. An introduction to decision tree modeling. J. Chemom. A J. Chemom. Soc. 2004, 18, 275–285. [CrossRef]
34. Vijayvargiya, A.; Prakash, C.; Kumar, R.; Bansal, S.; Tavares, J.M.R.S. Human knee abnormality detection from imbalanced sEMG data. Biomed. Signal Process. Control. 2021, 66, 102406. [CrossRef]
35. Vijayvargiya, A.; Kumar, R.; Dey, N.; Tavares, J.M.R. Comparative analysis of machine learning techniques for the classification of knee abnormality. In Proceedings of the 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 30–31 October 2020; pp. 1–6.