



OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG



FAKULTÄT FÜR
INFORMATIK

Otto-von-Guericke-University Magdeburg

Faculty of Computer Science

Department of Intelligent Cooperative Systems

Application of Genetic Algorithm-based
Feature Selection for Canine Musculoskeletal
Injury Identification

Bachelor Thesis

Author:

Jonas Seib

Examiner and Supervisor:

Prof. Dr.-Ing. habil. Sanaz Mostaghim

2nd Examiner:

Dr.-Ing. Christoph Steup

Magdeburg, June 11, 2024

Abstract

Musculoskeletal injuries in canines can significantly impact their well-being and quality of life. The detection of these orthopedic injuries is crucial for effective treatment. Despite the advancements in modern technology, this is not always feasible due to financial or practical limitations. This thesis examines the use of machine learning algorithms, including Logistic Regression (LR), k-Nearest Neighbors (KNN), and Support Vector Classifier (SVC), in order to determine which of these is best suited to identify an abnormal canine gait based on selected features. Random features are added as noise to assess whether classifiers identify and filter out irrelevant attributes from the dataset. In this thesis, the application of Genetic Algorithm-based (GA) feature selection is proposed in order to determine the most relevant features necessary for canine musculoskeletal injury detection and to increase the predictive performance of the classifiers. In the GA, every individual represents a unique subset of selected features. These subsets are subsequently employed to train the classifier and evaluate its predictive performance. The performances of the machine learning algorithms are evaluated using F1-scores and utilized as the fitness scores of the GA. Each classifier is assessed using the default and tuned hyperparameters. Results demonstrate the robustness and predictive efficiency of the SVC, as this classifier is able to filter out all random noise. The features most frequently selected as relevant to the target value are physical attributes of the dogs, including age, body condition score, weight, and sex. The paw area, average pressure, and their respective standard deviations, along with impulse, are chosen as the second most frequently selected features, particularly by the KNN and SVC.

Contents

1	Introduction	3
1.1	Motivation	3
1.2	Goals	4
1.3	Thesis Outline	4
2	Theoretical foundation	6
2.1	Optimization Methods in Data Analysis	6
2.1.1	Evolutionary Algorithms	6
2.1.2	Feature Selection	10
2.2	Canine biomechanics	12
2.3	Data Source	14
3	Related Work	16
3.1	Canine Gait Analysis	16
3.2	Feature Selection	19
4	EA-based Feature Selection on Canine Gait	22
4.1	Evolutionary Feature Selection Approach	22
4.2	Data Preparation	23
4.2.1	Visualization	23
4.2.2	Paw Detection	27
4.2.3	Features	28
5	Evaluation	31
5.1	Experiments	31
5.1.1	Preprocessing	31
5.1.2	Algorithms	32
5.1.3	Hyperparameter tuning	32
5.1.4	Feature Selection	33
5.2	Results	35
5.3	Limitations	41
6	Conclusion and Future Work	42

Bibliography	44
Selbstständigkeitserklärung	50
Statement of Authorship	51

1 Introduction

“To his dog, every man is Napoleon; hence the constant popularity of dogs.”

—Aldous Huxley [1]

1.1 Motivation

Although dogs are considered to be man’s best friend, we cannot always understand and comprehend their needs or ailments. This fact becomes even more apparent when an animal is in pain and even the best veterinarians are unable to determine the cause of a dog’s discomfort. Aside from obvious diagnoses such as superficial wounds or infections, some diseases are much more difficult to detect, even with the advancement of modern technology such as Computed Tomography (CT) or Magnetic Resonance Imaging (MRI), available not only for humans but also for animals. While these advanced imaging techniques allow for non-invasive diagnosis, they tend to present the challenge of handling a live animal in a confined space such as an MRI tube, while keeping the body part being examined as still as possible [2].

In many cases, the solution is to sedate the animal and add stress to an already sick or injured patient in order to obtain a medical diagnosis. Other drawbacks of these imaging technologies are the relatively high investment costs for equipment and maintenance, training and education of medical staff, as well as the regulatory nature and legal obligations when it comes to handling radioactive pharmaceuticals in general [3].

Furthermore, the limitations of some of these imaging practices should be considered, for example distortions of the images or disturbances caused by foreign metallic objects like pet identifying chips implanted in dogs and cats [4]. Since veterinarian organizations also have to be able to compete economically, acquiring the described types of equipment and machinery could be hindered by their high costs [5].

Clearly, these practices are predominantly used for more severe cases of illnesses and diseases in all sorts of animals. Beside disorders of skin, digestive tract, eyes, ears, and teeth, diseases of the musculoskeletal apparatus account for one of the prevalent reasons dog owners bring their canines to a veterinarian facility. The affected body parts include

muscles, joints, and bones and can be categorized in the orthopedic division of veterinary medicine and are responsible for about 8-17% of disorders veterinarians diagnose [6, 7]. The most common ones of these being fractures, arthritis, especially in older dogs [8], luxations, ligament ruptures and lameness. [9] states that 41% of the latter are of unclear diagnosis and as a result, clarifying the cause is not pursued any further. The negligence of clearing up the causes for lameness could then potentially lead to a decline in health condition in the future if not treated adequately.

1.2 Goals

The primary objective of this thesis is to investigate an alternative approach to diagnosing dogs with orthopedic problems using gait data and machine learning techniques. This study aims to address key research questions related to the effectiveness of feature selection and the predictive performance of different machine learning classifiers in identifying orthopedic issues in canines. The first research question investigates which machine learning classifier—among Logistic Regression, K-Nearest Neighbors, and Support Vector Classifier—exhibits the highest predictive performance for diagnosing ill dogs based on extracted features from the gait data. The second research question seeks to determine what conclusions about a canine’s orthopedic health can be drawn from feature selection in a canine movement analysis. Specifically, this involves identifying which features are deemed most relevant through the application of genetic algorithms for feature selection. The objective of this thesis is to contribute to the development of more accurate and efficient diagnostic tools for canine orthopedic health, with the ultimate goal of enhancing the quality of veterinary care. Furthermore, this thesis aims to increase insight into the predictive accuracy of detecting ill dogs using machine learning algorithms.

1.3 Thesis Outline

Chapter 2 establishes a foundation of all the relevant theoretical groundwork regarding the topic of feature selection with the help of evolutionary algorithms. Furthermore, it addresses the data source and acquisition process. **Chapter 3** discusses related works regarding the topic of feature selection and evolutionary algorithms, as well as the theoretical foundation on canine biomechanics and movement dynamics. Moreover, the chapter reviews related works on the topic of canine gait analysis using different technical devices. The structure and execution of the conducted experiments are explained in depth in **Chapter 4**. The results and findings of these experiments are presented and evaluated

in Chapter 5.

Finally, Chapter 6 will summarize the thesis and give an outlook for possible future work.

2 Theoretical foundation

This chapter provides an examination of the foundational knowledge essential for this thesis. It encompasses optimization techniques such as evolutionary algorithms and feature selection, as well as fundamental terminology related to canine gait. Additionally, it offers a concise overview of the data source utilized for the experiments.

2.1 Optimization Methods in Data Analysis

This chapter examines the domain of optimization methods in data analysis, with a particular focus on evolutionary algorithms and feature selection techniques. Evolutionary algorithms offer a computational approach, inspired by natural selection, to explore solution spaces to identify optimal solutions in complex optimization problems. Additionally, feature selection is presented, a data preprocessing technique that aims to identify and retain the most informative features while discarding redundant or irrelevant ones.

2.1.1 Evolutionary Algorithms

In the following, a brief overview and description of Evolutionary Algorithms and genetic algorithms in particular will be given. Then, the characteristic components of GAs will be presented in detail.

Evolutionary Algorithms (EAs) are a set of computational optimization techniques that simulate the process of natural evolution, as described by Darwinian theory. This class includes Evolution Strategies, Evolutionary Programming, and Genetic Algorithms (GAs). GAs were first introduced by Holland in his work 'Adaptation in Natural and Artificial Systems' in 1975. They are used to solve optimization and search problems where the optimal solution is unknown and exploring the entire solution space is infeasible. GAs represent each potential solution as a chromosome and assess the quality of each solution using a problem-specific fitness function. GAs iteratively improve the population

through iterative generations of selection, reproduction, and evaluation, converging towards optimal or near-optimal solutions for the given problem. Due to their parallelism and universality, GAs are often superior to traditional algorithms when dealing with complex problems [10].

Genetic algorithms use terminology derived from biology to describe their processes. A GA begins with an initial population of potential solutions, referred to as 'chromosomes' or 'individuals.' These are typically represented as binary strings, although other data structures, such as real-valued vectors, are also possible. Initialization of the first generation is often random to avoid any bias in the beginning. By incorporating a random component into the GA, some solutions may fall outside the space of valid solutions for the specific problem domain. It is up to the user to define how the genes of a chromosome translate to the specific problem domain, as the algorithm itself cannot achieve this. A chromosome is composed of multiple 'genes', represented as 0 or 1 in binary encoding. Each gene can be considered a feature of the particular solution with unique characteristics, depending on the solution space. The fitness value of each chromosome is evaluated by a fitness function in each iteration or generation of the algorithm. The selection mechanism gives a higher chance of reproduction to solutions that are the fittest in their respective generation with respect to the specific optimization objective(s). A fitness function assesses the individual fitness value. The chosen individuals then become parents for the next generation. Depending on the algorithm's implementation, parent individuals may also persist in the next generation. This process simulates natural selection by favoring the reproduction and survival of the fittest. Offspring are created by applying genetic operators, such as crossover and mutation, to two parent chromosomes. Crossover rearranges blocks of genes from the parents to create offspring with genes from both parent chromosomes. In that way, new combinations of solutions are generated and that region of the search space is exploited. To ensure variation and diversity in the solutions, the mutation operator randomly flips genes in a chromosome with a low probability. This guarantees exploration of the search space. However, GAs perform well in finding a balanced trade-off between exploring promising solution regions and exploring the whole space, which tend to be competing objectives. This process is repeated until a satisfactory solution is found, or termination criteria are met. Termination criteria may be based on computational hardware or time limitations, or the maximum number of generations reached. In the final generation, a set of potential solutions is presented, making GAs a robust approach to optimization problems compared to single-solution optimization methods [11].

When implementing a genetic algorithm (GA), it is necessary to set parameters which vary greatly depending on the specific problem being optimized. These parameters include probabilities for crossover and mutation, as well as population size. Although generation

size as a termination criterion can impact the performance of the GA, it is not further examined here. It is important to note that no generalization for the values of the parameters can be drawn, since they highly influence each other. Therefore, the parameters cannot be chosen arbitrarily, but rather must be optimized with respect to the overarching objective(s). The probability of crossover affects how likely it is for chromosomes to undergo the genetic crossover operator. The mutation probability sets the rate at which a gene is subject to random alteration. In binary encoded chromosomes, this translates to a simple bit flip [12].

There are two approaches to optimization problems: single-objective and multi-objective optimization. Single-objective optimization aims to optimize a single criterion, such as minimizing the total distance traveled in a route planning algorithm. Since the goal is to optimize a single objective, the problem is called a global optimization problem. The fitness function quantifies how well a proposed solution performs in regard to that single objective goal.

Multi-objective optimization involves multiple (often conflicting) objective functions, such as maximizing fuel efficiency while minimizing costs in car manufacturing. The goal of optimization is to find a set of solutions that provide a trade-off, where no objective can be increased without worsening the other. These solutions are known as Pareto optimal solutions. Fitness functions in multi-objective optimization problems incorporate a metric to compare solutions based on their performance across multiple objectives. Therefore, the metric may indicate similar fitness values for multiple potential solutions, even if the objective functions in the solutions differ [13].

The first step of a GA is to create the initial population. This is often done randomly to avoid any bias an informed or predetermined starting population could introduce. Then, each chromosome is evaluated and the fittest ones with the highest fitness value are selected for reproduction. The selection of the most fit individuals is performed by selection operators. One of the most common selection techniques is the tournament selection, which randomly chooses some n chromosomes from the population and selects some number of the fittest chromosomes. There also exist other techniques, like the roulette-wheel selection, which favors individuals with relatively higher fitness than the population average. It assigns each of the chosen n chromosomes their proportional share of the roulette-wheel to the ratio of the individual's fitness and the average fitness of the population. The choice of the selection technique has a great impact on the algorithm's performance, since a trade-off between exploration and exploitation happens [13].

The selection techniques employed do not guarantee the retention of the optimal solution in the subsequent generation. Consequently, 'elitism' can be implemented. This mechanism entails the preservation of the optimal solutions, regardless of the outcome of the selection algorithm [10].

Subsequently, genetic operators such as crossover or mutation are applied to the selected

chromosomes. One of the fundamental crossover operators is single-point crossover, which represents a form of recombination between two parent chromosomes. The two parent chromosomes are randomly cut at a specific gene, after which the cut-out portion of one parent is recombined with the other parent. Mutation operators in binary-encoded problems are relatively straightforward to perform, as they entail a single bit flip of a randomly selected gene in an individual. Another mutation operator is the switching of two different genes in a single individual [14].

The evaluation of an individuals' fitness can be done by a metric or function that states the goodness of the presented solution. A higher fitness therefore means higher chances of reproduction for that chromosome. In many cases, the fitness function of single-objective optimization problems can be directly derived from the objective function, since it represents the quantifiable optimization goal. In multi-objective problems, the fitness function has to factor in multiple objective functions to quantify how well a presented solution accomplishes those goals [13].

In optimization problems involving classification, the goal is to create a model to accurately categorize input data into predefined classes or categories based on the features of the data. During model training, the model is given input variables and their associated target. After model training, the performance is evaluated on a separate test dataset using evaluation metrics such as accuracy, precision, recall, or F1 score.

In binary classification problems, there are four possible outcomes of classification: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). TP and TN occur when the model correctly predicts a positive or negative instance, respectively. FP and FN occur when the model incorrectly classifies an instance as the opposite case.

Accuracy measures the proportion of correctly classified instances out of the total instances in the dataset. It is a commonly used metric for evaluating classification models.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Precision measures the proportion of true positive predictions out of all positive predictions made by the model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall measures the proportion of true positive predictions out of all actual positive instances in the dataset.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1 score is the harmonic mean of precision and recall and one of the most commonly used

classification metric. It provides a single metric that balances both precision and recall, making it useful for evaluating models when there is an imbalance between positive and negative instances [15].

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Model training and evaluation is repeated until the metric reaches a satisfying result. Then, the model can be used to predict class labels of new, unseen data based on the learned patterns and relationships [16].

2.1.2 Feature Selection

The exponential growth of data presents both opportunities and challenges for data-driven industries. While the abundance of data promises valuable insights and opportunities for innovation, it also poses significant challenges, particularly in managing and extracting meaningful information from vast and complex datasets. As the volume of data increases, so too does the computational burden associated with processing, analyzing, and interpreting it. Furthermore, the presence of redundant, irrelevant, or noisy data serves to aggravate these challenges, with the potential consequence of decreasing model performance and decision-making [15].

Feature selection is a process in machine learning and data analysis whereby the most relevant features or variables are identified from a larger set of available features. The objective of this technique is to enhance the performance of the model, prevent overfitting, and facilitate interpretability by selecting a subset of features that contribute the most to the predictive power of the model. At the same time, it aims to discard irrelevant or redundant ones which provide no additional value in solving the problem at hand. Feature selection is of value for a number of reasons. Firstly, it helps to simplify models by reducing the dimensionality of the input space, which can lead to faster training times and lower computational costs. Furthermore, it enhances the model's generalization capacity by concentrating on the most informative features, thereby reducing the probability of overfitting to noisy or irrelevant data. Moreover, feature selection enhances the interpretability of the model by highlighting the key factors driving the predictions [17].

However, feature selection presents its own set of challenges, particularly in large-scale datasets, where the combinatorial explosion of feature subsets makes the task of identifying the optimal subset computationally expensive. Determining which features are truly useful and which can be safely discarded without sacrificing predictive performance is an NP-hard problem. The issue becomes even more complex when considering that certain feature subsets are more effective for specific classifiers [18, 19].

Feature selection encompasses three main approaches: filter methods, wrappers, and embedded methods. Filter methods assess the relevance of features directly from the underlying data and independently of the chosen model. These techniques evaluate features based on statistical measures or tests, such as correlation coefficients, information gain, or relevance, in order to rank or select them by their importance. The main advantage of the filter approach is that it is computationally less expensive than the other approaches. The issue with this approach is that redundant features cannot be discarded in this manner, as they are ranked in the same order. When only one feature is being discarded at a time, the other redundant features that might also rank the same remain in the feature subset [20].

The wrapper approach entails integrating feature selection directly into the model evaluation process. In contrast to filter methods, which assess features independently of the selected model, wrappers evaluate subsets of features based on their performance within the context of a specific model. This iterative process typically involves selecting a subset of features, training the model using only those features, and evaluating the performance of the model based on a predefined metric, such as accuracy or precision. The subset of features that maximizes the performance metric is retained, and the process continues until a stopping criterion is met. Wrappers provide a more accurate assessment of feature importance compared to filter methods, but they are more computationally expensive, as they require training and evaluating the model multiple times for different feature subsets. Evaluation of the model can include multiple objectives like a predefined metric of the model as well as the number of chosen features. Here, a higher score on the metric and a lower number of features would be an ideal solution [19].

The search for a subset can be conducted in a naive manner, analogous to the greedy forward search. In this approach, the subset is initiated with a randomly selected feature from the total subset. Subsequently, another feature is incorporated into the subset in each iteration, and the predictive capabilities of the model are evaluated. This process is repeated until the next feature added results in a decrease in performance. The reverse counterpart of this method is called backward elimination. It begins with all features included in the subset, and then iteratively removes a random feature until the stopping criterion is met. In the worst case, the complexity of these heuristics degrade by a factor of $O(n^2)$ with n being the number of features [19].

Given that the feature subsets can be described as strings of 0s and 1s, with the value indicating whether the feature is included in the subset, it is possible to incorporate more sophisticated search heuristics like a combination with a GA. Firstly, the modification of the feature subset string is of great interest. During the alteration of the subset string, the convergence of the solution may be accelerated if multiple features are added or disregarded in a single iteration, compared to forward selection or backward elimination described earlier. Secondly, it is also important to consider the termination criterion.

This could be, for example, until no improvement in performance is observed with the addition of features, until the search space is exhausted, or until some runtime limitation is reached [21].

Embedded methods integrate elements of both filter and wrapper methods. Incorporating feature selection within the model training process, these methods select features based on their contribution to model performance during training. Techniques such as Lasso regression and tree-based algorithms implement embedded feature selection by imposing penalties or pruning irrelevant features. This enables the simultaneous training and selection of models, thereby leading to more efficient and accurate models. One example for embedded methods is the Recursive Feature Elimination for Support Vector Machines, which iteratively trains an SVM classifier and keeps only the most important features based on the weights of the SVM solution [15, 20].

2.2 Canine biomechanics

Canine gait and biomechanics are essential aspects of understanding how dogs move and function. By examining the sequences of leg movements and the underlying biomechanical principles, insights into their locomotion capabilities can be gained and a theoretical knowledge foundation is established.

Gait is defined as a 'series of repetitive motions' [22] consisting of multiple strides. Each stride has a distinct step cycle each limb performs. A step cycle comprises a stance phase and a swing phase. The stance phase is the contact time between paw and ground, whereas the swing phase is the time of the paw being airborne.

Different canine gaits are characterized by distinct patterns of leg movements, known as paw sequences. In the following, the paw designations shall be abbreviated as follows: front right - FR, front left - FL, back right - BR, back left - BL. These sequences, such as RF, LH, LF, RH for the 'walk', correspond to specific combinations of leg support during movement. To allow for controlled movement, especially at slower speeds, multiple paws are in contact with the ground at the same time. If that is the case, the set of paws in stance phase, i.e., touching the ground, are called 'supports' [23]. Dogs typically transition between these support patterns, ensuring that ideally three paws are on the ground for maximum stability. These gait patterns, including walk, pace, and trot, serve various purposes, from conserving energy during walks to achieving faster speeds.

The walk is the most energy conserving gait, with at least two paws on the ground all the time and three or four paws at some times. The center of gravity in a slow walk is maintained somewhere within the triangle that three of the paws form, enabling the fourth paw to move freely without tipping the rest of the body. When accelerating, the center of gravity cannot be maintained within that triangle and shifts towards the free

paw. This causes the rest of the body to 'fall' onto that side of the fourth paw, forcing it to cushion the landing [23].

At that point, a lot of dogs alter their gait to the pace, which is slightly faster than the walk but less tiring as the legs on the same side move forward almost simultaneously. Both the walk and pace are symmetrical gaits, since each limb pair copies the other side in the following movement. One downside of the pace is the missing support of the opposite side during a step. For instance, when both FL and BL touch the ground, center of gravity causes the dog to lean towards its right side. Then, FR and BR move forward to counterbalance falling to the side, resulting in left-sided tilting. As a consequence, the dog sways sideways, which is energetically suboptimal [23]. The further apart the legs are, the more the dog sways due to a longer leverage. This gait can be slow with momentarily all four paws on the ground, or faster with a brief flight phase.

Another gait is the trot, where diagonally opposite limbs, e.g., (FR+BL) or (FL+BR), are moved at once, which is less tiring than the pace but also slower. This gait is mostly used in orthopedic evaluations as it is symmetric and a planted leg carries weight without support of the opposite leg [22]. Thus, instabilities or asymmetries, which can be indicators of muscular or orthopedic conditions, are easier to detect [22].

Causes for disruption of the movement apparatus, like lameness, can arise from limited range of motion of joints or injuries of tissue involved in the movement. Other factors that have an impact on the quality of canine movement are surface, incline, and curvature of the pathway. In most healthy dogs, the front limbs typically bear approximately 60% of the body weight when walking on a horizontal surface, while the hind legs carry 40% [22]. When walking uphill, that allocation shifts more towards the hind legs, which then take over as the main propulsion force. That is one of the main reasons why gait analysis is most significant when performed on an even ground. Treadmills are considered a special case in canine gait analysis as the ground moves beneath the center of gravity as opposed to 'normal' overground movement where it is vice versa [22]. Furthermore, [24] states that horses show significant differences in movement compared to normal overground motion. As this phenomenon is not yet investigated thoroughly in dogs, this area requires further research in the future.

For many years, dogs have lived alongside humans, allowing humans to influence the development of dogs. Different dog breeds exhibit unique biomechanical adaptations suited to their environment or working purposes. For example, breeds in northern regions may have insulating fur to withstand cold temperatures, while retriever dogs feature webbing between their toes for efficient swimming. As a result, different dog breeds vary in conformation. This includes different body characteristics and forms relative to other parts of the body but also shape and symmetry. Some breeds may even possess a pleasing gait style reminiscent of horses, or develop gait abnormalities in order to compensate for anatomic abnormalities. An example would be a skewing gait for dogs with long hind

limbs. When a hind leg swings forward to take a step, it is possible that the hind leg might touch the front leg on the same side when the front leg does not move fast enough. To not impede its own gait, some dogs walk slightly sideways to allow the hind leg to reach next to the front leg [23].

Modern techniques provide detailed insights into canine locomotion. There are two distinct categories of gait assessment techniques, kinematic and kinetic analysis. Kinematic analysis 'quantifies the positions, velocities, accelerations, and angles of anatomic point, segments, and joints' [22], i.e., without directly including forces in the evaluation. Kinetic analysis evaluates forces, impulses, and pressure distributions exerted during movement [24]. These techniques allow researchers to study subtle variations in gait and understand how different factors impact canine locomotion. For kinematics, variations in anatomy, skin movements and other uncertainties can be introduced, resulting in limitations of data interpretation. In kinetic analysis, factors like surface or material of the walkway can have an impact on the measurement [24]. Disadvantages of force plate-based kinetic analysis are the limited stride length-measuring capabilities, a complex study setup and evaluation, and high costs [25]. Some of these detriments are eliminated by pressure sensing techniques, which on the contrary come with their own disadvantages, for instance the loss of force dimension separation [25].

In canine rehabilitation, gait analysis plays a crucial role in assessing movement patterns and identifying abnormalities. By analyzing repetitive motions and evaluating the stance and swing phases of each leg's step cycle, clinicians can detect issues such as lameness or joint dysfunction. Symmetry in gait is particularly important for maintaining a healthy locomotor system and overall well-being. But nonrepetitive motions too can give valuable insight into an animal's health status. Motions like the start of a movement or jumping are not only executed with a bigger range of motion but also with some of the highest forces throughout the entire movement [22].

Visual gait analysis performed by veterinarians is not able to discern these nuanced details as reliable as with the help of modern technologies [25]. Therefore, development of diagnosis systems aiding physicians is imperative to enhance diagnosis quality in the future.

2.3 Data Source

The measurements of the dogs were performed by the [University of Veterinary Medicine Hannover](#) (TiHo) since the project is a cooperation between the Otto-von-Guericke University Magdeburg (OvGU) and the TiHo. The study uses two different groups of dogs for data collection. The first group consists of healthy dogs kept at the TiHo as clinic dogs. The second group consists of injured or ill dogs that are treated as patients at the

TiHo clinic. Their owners have voluntarily agreed to have their dog's gait measured after receiving treatment. During the measurements, the owners accompany their dogs, but under the supervision and guidance of TiHo staff, while clinic dogs are led by TiHo staff. The data is measured by a 200 Hz pressure sensor mat that the dogs walk across from one end to the other while being led on a leash. When the dog leaves the end of the mat, that particular gait is completed. The handler then turns the dog and themselves around, preparing the dog for the next walk in the opposite direction. The handler never touches the mat and walks along the exact same side of the mat for each walk. This, combined with the alternating starting side of the mat, ensures that any bias in the data caused by the dog leaning to the handler's side of the mat is evened out over multiple gaits.

Each dog is measured twice, with one measurement being as slow as possible without allowing the dog to come to a complete stop ("walk"), and the other being a fast trot within the dog's range of motion, however, no running is allowed. Each of these two measurements consists of several walks, which averages out to 5.57 walks per dog for the trot. The scope of this thesis focuses solely on the trot dataset, as it was found to be the more reliable and larger in terms of the total number of walks measured.

The trot dataset consists of 197 dogs with a total of 1097 measured walks. Of these 197 dogs, 52 healthy dogs performed a total of 306 walks and 145 sick or injured dogs performed a total of 791 walks. To filter the data set for faulty walks, each walk is considered as a single data point and validated to ensure the reliability of that particular walk. An in-depth explanation on data validation can be found in [Section 4.2](#). Each gait is then labeled with the corresponding medical status of the measured dog.

The number of walks performed in both directions may vary for each dog, depending on factors such as gait quality, cooperation, and the patient's level of pain. Handling of potentially erroneous data is done later in the data preparation process, as explained in [Section 4.2](#). Importantly, the measurement process remains consistent across both groups.

Additional patient information is also recorded, including height, weight, Body Condition Score as a measure of obesity, breed, sex, spay/neuter status, pain level, and any diagnosed medical conditions, if available, and linked to the dog's corresponding sensor pressure data.

3 Related Work

This chapter explores two key areas that this thesis combines, namely canine gait analysis and feature selection, with a focus on evolutionary algorithms for building predictive classification models.

3.1 Canine Gait Analysis

The detection of musculoskeletal diseases in dogs can often be subjective and biased, not only by experienced veterinarians, but even more so by dog handlers, who may not recognize a potential disease until the animal shows clear signs of discomfort, such as limping, lameness, or tenderness in the affected limb. At this point, other negative, potentially irreversible health effects may occur, resulting in further deterioration of health. Additionally, the treatment of the advanced disease may prove to be more difficult at a later stage comparatively to treatment at an earlier stage of development.

To address the earliest possible detection of orthopedic conditions, many studies have investigated the analysis of canine gait by either focusing on the pressure exerted by the paws on special force/pressure sensing devices, such as force plates or pressure mats [26–29], or by looking at the angular movement of limbs and joints during locomotion [30, 31]. The reviewed existing work can be classified in multiple categories. The first category includes studies which examine healthy dogs of the same or different breeds and compare them in their analysis. Same-breed studies provide baseline data for that specific breed [26, 32]. Whereas multi-breed studies compare breed-optimal movement patterns between different breeds. To what extent they are comparable and what variables influence these comparisons is based on healthy individuals [33–35].

[26] studies the back motion of German Shepherds (GS), combining both angular motion captured by multiple cameras with reflective markers at key anatomical locations, and force data measured by force platforms integrated into the surface of the walkway floor. In contrast to the method proposed in this thesis, where step detection is automated, the previous study uses manual step detection by marking the matching time frame in the video to align the video data, and thus the 3D motion, with the measured force data.

[32] studies only the left limbs of healthy greyhounds using two force plates two meters

apart to evaluate the variance in the forces and impulses recorded, as well as the influence of different handlers guiding the dogs throughout the measurement. Since the handlers account for only 0-7% of the variance in the data collected [32], the influence of different handlers is ignored in the underlying data set of this thesis.

While [33] concludes that GS and Labrador Retriever (LR) dogs have major differences in their 2D angular body model derived from bone measurements based on anatomical markers, limb weight-bearing and gait parameters are mostly similar after normalizing for dog size and weight. Moreover, the paper states that there appear to be genetic predispositions that result in different anatomical structures, such as the center of pressure being closer to the head in LR dogs than in GS dogs.

In contrast, [36] analyzed seven different breeds with at least five healthy individuals each. Despite normalizing peak vertical forces, vertical impulses, stance times, and the ratio of forelimb impulse to total impulse, they concluded that comparison may not always be recommended due to fundamentally different body conformations, one exception being the case of reference groups that share body dimensions such as height, limb length, and weight.

Another study has come to a similar conclusion that the measured gait data in dogs cannot be compared as easily as in other species of animals due to the wide variety of canine body conformations [27]. According to dynamic similarity, a principle which states that animals of the same species can have similar movement characteristics regardless of their size ratios, gait data is mostly similar when biological and morphological differences such as relative preferred speed, stride frequency and length, and body mass are accounted for [37]. Therefore, studies in this field of work use ratios between limbs [35, 38], similarity indices [28, 39], or normalization for various earlier mentioned morphological variables [29, 40].

This section pertains to studies that investigate dogs with musculoskeletal disorders such as hip dysplasia (HD), osteoarthritis, or ligament ruptures. The studies aim to derive non-invasive and accurate methods of identifying and diagnosing these diseases using various techniques [28, 31, 41]. These include angular-based, pressure-/force-based methods, or a combination of both [29].

Studies that utilize pressure walkway systems to detect lameness have found that temporospatial and kinetic variables, such as peak vertical force (PVF), vertical impulse (VI), and paw contact area (PCA), measured on a pressure sensor walkway or plate are effective in identifying a lame limb. These variables tend to be significantly lower in the lame limb compared to the non-lame limb [28, 42].

Symmetry indices or ratios comparing same-side or lateral-side limb pairs are effective in identifying lameness. In this case, pressure sensor walkways are favored instead of force plates, as the former is able to measure load on the paw with a higher resolution and across a wider range of the walkway, resulting in valuable information about a paw's

weight-bearing change over time and distance [42].

Similar to the method proposed in this thesis, gait variables such as stride length or PCA were calculated based on the distance between sensor cells or the number of activated sensor cells, respectively.

Multiple studies incorporate some form of video recording of the trials, either to detect invalid trials based on predefined exclusion rules [41] or to analyze angular limb and joint movement on video [30] or in a 3D model of the dog [31]. Capturing the measurements on video for validation purposes could prove as a great overhead, considering acquisition, setup, and calibration of the equipment, and evaluation of video data. This could limit the practicability of using video as a diagnostic system in real-world applications. While 2D videos are easier to handle and cheaper, 3D models of the research participants provide more in-depth information about the subject at hand [43].

In a carefully arranged study setting, it is easier to ensure lighting conditions and detectability of attached markers on pivot anatomical locations compared to real-world diagnosis environments.

Furthermore, reflective markers are attached to the skin or fur, causing imprecision due to skin movement, which can affect accurate joint and bone monitoring [31].

One study incorporates both 3D video analysis and force plate data to evaluate methods of HD treatment. However, the results indicate that kinematic data and force plate data, e.g., peak force or stride length, alone could potentially be sufficient in detecting dogs with HD [41]. While omitting the 3D video analysis would remove the visual component of the evaluation, it would also reduce the overall study overhead and complexity.

One alternative method for evaluating canine gait patterns is through the use of small wearable motion sensors attached to the dog's limbs [40, 44]. These sensors can automatically detect single steps performed by a specific leg and distinguish between different gait patterns, such as walking or trotting. Those devices are not only affordable but also capable of accurately detecting the exact moments of placing and lifting of the paw based on angular velocity and acceleration of the gyroscope signal of the wearable with an accuracy of 99.92% [40]. However, the authors of the study concede that limb-mounted wearables greatly impair the walking pattern greatly initially after attachment. They affirm that after a short period of habituation, all canines revert to their previous gait style [44]. Additional research is necessary to determine if wearables could potentially bias stride in an imperceptible way.

In summary, when evaluating canine gaits, both angular and kinematic strategies provide objective and scientifically sound data. However, angular strategies may not be as feasible in real-world scenarios. Pressure- or force-based technologies may require normalization or the use of ratios or indices to compare data between multiple individuals, as some studies have concluded. Pressure sensing plates are becoming more commonly used in recent studies because they provide insights into how force is distributed over time and also

offer more spatial information than force plates. However, force plates have the advantage of being able to measure additional data, such as center of pressure or force vectors, in addition to vertical forces. Wearable inertial sensor devices have the potential to enhance gait analysis and could serve as a supplementary or even a substitute technology for well-established sensors.

3.2 Feature Selection

This section will examine existing works on various feature selection (FS) methods, with a particular focus on those that integrate evolutionary algorithm techniques. FS is a pre-processing technique used in data mining and machine learning applications to include only relevant features in the main processing stage, with respect to the overall data mining or machine learning objective. Determining whether a feature is relevant to the target and belongs to the optimal feature subset is an NP-hard problem [15].

FS is a crucial step in data analysis that holds immense value across diverse domains. FS involves selecting the most relevant features from a dataset while discarding irrelevant or redundant ones. This not only enhances model performance but reduces computational complexity. This technique finds applications in a wide array of fields, such as credit rating [45], iron ore quality monitoring [46], face recognition [47, 48], spam detection [49], video data classification [50] and bioinformatics [51].

[52] employs a correlation-based filter approach to identify 'good' features for the optimal feature subset. A feature is considered 'good' if it exhibits high correlation with the target value/class, but not with other features. The paper cited above measures correlation using symmetrical uncertainty, which quantifies the shared amount of information between two variables, and a proposed concept called 'predominant correlation' that integrates redundancy between two relevant variables.

[17] presents various statistical methods for filtering relevant features in their work. The present features can be ranked using criteria such as R^2 , Pearson correlation, or mutual information, and the best performing features are used for model training.

Filter methods are computationally faster than wrapper methods and less prone to overfitting. However, they do not consider the classifier used during the training phase and might overlook possible feature interactions. It is possible for a feature that seems irrelevant based on some ranking criteria to become relevant when taken with one or more other features. The same applies to two irrelevant (ranked) features which in combination might be useful [17].

Combining filter and wrapper approaches can help avoid missing presumed irrelevant features. This can be achieved by prefiltering and selecting the final feature subset through

following wrappers, as demonstrated by several studies [45, 51, 53, 54].

[51] utilized the F-score and information gain methods before employing wrapper methods, specifically a combination of backward elimination and forward selection. To estimate the discrimination ability of feature subsets, an SVM was trained simultaneously until accuracy decreased. The best performing subset was then determined to be the overall best solution.

In [53], a comparison was made between a simple genetic algorithm (GA), a hybrid filter-GA method, a forward search, and a backward elimination on both artificial and public-domain datasets. The hybrid method used a class separability metric as the filter and a GA as the wrapper. The GA's fitness was determined by a Naive Bayes' estimate of accuracy on the generated feature subset. A different study employs a combination of measures for both information content of each feature and its extraction cost and classification accuracy of a decision tree classifier [54].

[45] utilizes information gain to prefilter the initial feature set and propagates the best ranked features to a GA wrapper. The GA wrapper employs three different classifiers: K-nearest neighbor (KNN), Naive Bayes, and SVM. These classifiers are tested with 10-fold cross-validation, and the averages are presented.

Although the hybrid approach performs similarly or better than the straightforward wrapper approach in terms of computation time and accuracy on larger data sets, it may negatively impact performance on smaller or lower dimensional data sets by disregarding features that are supposed to be irrelevant during the filter stage.

Apart from the hybrid method, some work focuses on the wrapper as the sole feature selecting meta-heuristic. One study, trying to enhance performance in facial recognition systems, uses a GA with KNN as chromosome fitness evaluation after image processing and feature extraction [48].

Thus far, only single-objective feature selection studies have been considered. However, some studies address multi-objective feature selection, where minimizing feature subset size is just as important as maximizing classification accuracy. One study utilized a fitness function that considers both classification accuracy and the number of features in the subset. Additionally, they used a mutual information metric to compare the selected features and the proportion of selected features in each solution [47]. A study on multi-objective feature selection utilized the NSGA-II algorithm to obtain a non-dominated feature subset. The selected features were then evaluated using various multi-label classifiers on video data [50]. The objective was to maximize prediction accuracy, as measured by different multi-label classifiers, while minimizing the number of features used.

The multi-objective approach can reduce the size of feature subsets while maintaining high prediction accuracy. However, it often requires problem domain-oriented knowledge to weigh objective factors or thresholds, and it has a higher computational cost.

In this thesis, a GA-based wrapper approach is applied to the feature selection problem of canine kinetic data measured on a pressure sensing walkway. This study is unique because it combines the feature selection problem in the domain of canine gait analysis.

4 EA-based Feature Selection on Canine Gait

This chapter explains the evolutionary algorithm-based feature selection approach. Furthermore, it covers the data preparation process, which includes data visualization and paw detection as the foundation for the subsequent feature creation, and the feature construction itself.

A GitHub repository of the code base can be found [here](#)¹. The code is written in the Python programming language (version 3.8). Additionally, the scikit-learn and sklearn-genetic-opt libraries were used for the experiments.

4.1 Evolutionary Feature Selection Approach

In order to address the research questions, the dataset must first be prepared and pre-processed. This involves the detection of the paws from the raw data, which is necessary for creating features associated with the paws. The created dataset consists of these extracted features, along with the dogs' medical records and body measurements. Following this, experiments are conducted using a genetic algorithm (GA) for feature selection. The GA searches for the minimal subset of features needed to predict a sick dog using specific machine learning classifiers. The classifiers employed are Logistic Regression (LR), K-Nearest Neighbors (KNN), and Support Vector Classifier (SVC). Each classifier is run twice with the GA, once with the default hyperparameters from the scikit-library and once with tuned hyperparameters using a partition of the data. This is done for two reasons. Firstly, this method allows the impact of feature selection on the prediction performance of default classifiers to be observed. Secondly, it is possible to achieve a higher overall performance when combining feature selection with hyperparameter tuning [55]. In order to ascertain whether the GA is capable of identifying and discarding irrelevant features from the subset, random features are incorporated into the data set as noise. The classifier's performance is employed as a metric to assess the fitness of the

¹https://github.com/nonajet/BA_Jonas_Seib.git

feature subsets proposed by the GA. The change in predictive performance between different proposed feature subsets is evaluated by comparing the classifier’s performance when trained solely on that specific subset. The GA is executed multiple times with identical hyperparameters and settings for each classifier, thereby enhancing statistical significance and preventing local optima from exerting too much influence on the final outcome. By analyzing which features are repeatedly selected for the minimal subset, conclusions can be drawn about the most important or overall relevant features, contributing to a higher accuracy. Another area of interest is the final predictive capability of the classifier when using the identified minimal subset in combination with optimized hyperparameters. This could provide insight into the introductory question of which classifier is most effective at recognizing a sick dog based on its gait and medical features. This analysis facilitates the identification of the critical features essential for the accurate prediction of a sick dog.

4.2 Data Preparation

In this thesis, all the runs of each dog during a single measurement are organized into a single XML file. This file follows a specific structure, as illustrated in [Figure 4.1](#). Each movement, which corresponds to a walk from one side of the mat to the other, is uniquely identified by a gait ID. The raw data captured consists of pressure values measured by the pressure sensors embedded in the mat. These values are recorded only in areas where pressure, such as from the dog’s footsteps, is detected. Additionally, for each matrix representing one two-hundredth of a second, the starting point on the entire mat is saved as an offset relative to a fixed corner of the mat. This allows for precise tracking and analysis of the movements of the dogs during the measurement process.

4.2.1 Visualization

The process of data preparation finds its foundation in the visualization of the underlying data to enhance understanding, validate the quality of the given data, and produce ideas on which potential features could be extracted from the gait analyses. The explanation about the feature structure is covered in-depth in [4.2.3](#). The visualization consists of three different views, which all present the data in different contexts in terms of time or location of the data.

Firstly, there is the visualization of the matrix data (‘local’) which is the raw data and can be considered an excerpt from the overall mat, depending on where the pressure sensors quantify pressure being exerted. As shown in [Figure 4.2](#), the distinct single paws

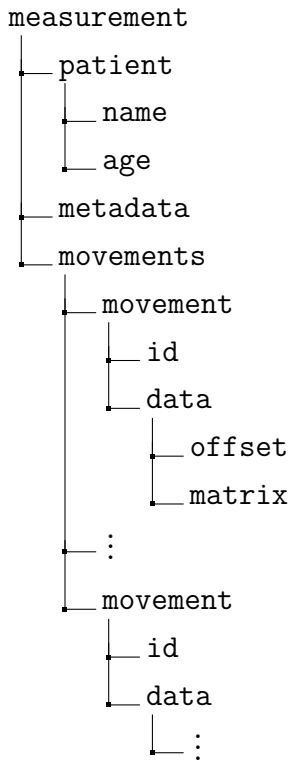


Figure 4.1: XML data file structure

are visible in the form of imprints on the violet background that is the part of the mat unaffected or not touched by the dog, respectively. The color differences between the paw imprints on the local view and their respective enlargement originate from the fact that each pixel is automatically colored according to the sensor pressure cell value (in N/cm^2) relative to the other pressure values present. Hence, resulting in few pixels in the local view potentially dominating the color spectrum, like in [Figure 4.2](#) happening for the back left paw, clearly visible with the single yellow pixel on the pad of the paw. This color domination is mitigated when each paw is individually visualized like in the enlargements, where smaller differences in pressure data for a single paw are more perceptible.

Secondly, a local matrix can, when combined with the according offset for the positioning on the whole mat, be embedded in a global view ([Figure 4.3](#)) which, over time observed, represents the absolute position of the dog on the mat at a given instant. Here, the color spectrum is also determined by the greatest relative pressure value measured.

Lastly, the total view stacks all global views from one gait on top of each other to provide insight into the pathway taken and other factors like rectilinearity of the walk or to what extent hind paws follow in the steps of the front paws. Towards the end of the visualization of a given walk, this view bears a resemblance to paw prints left in the snow.

Since the visualization does not offer confident and reliable distinction between the paws, a paw detection mechanism was implemented. That mechanism serves as the groundwork for the later computation of the different features extracted from the data.

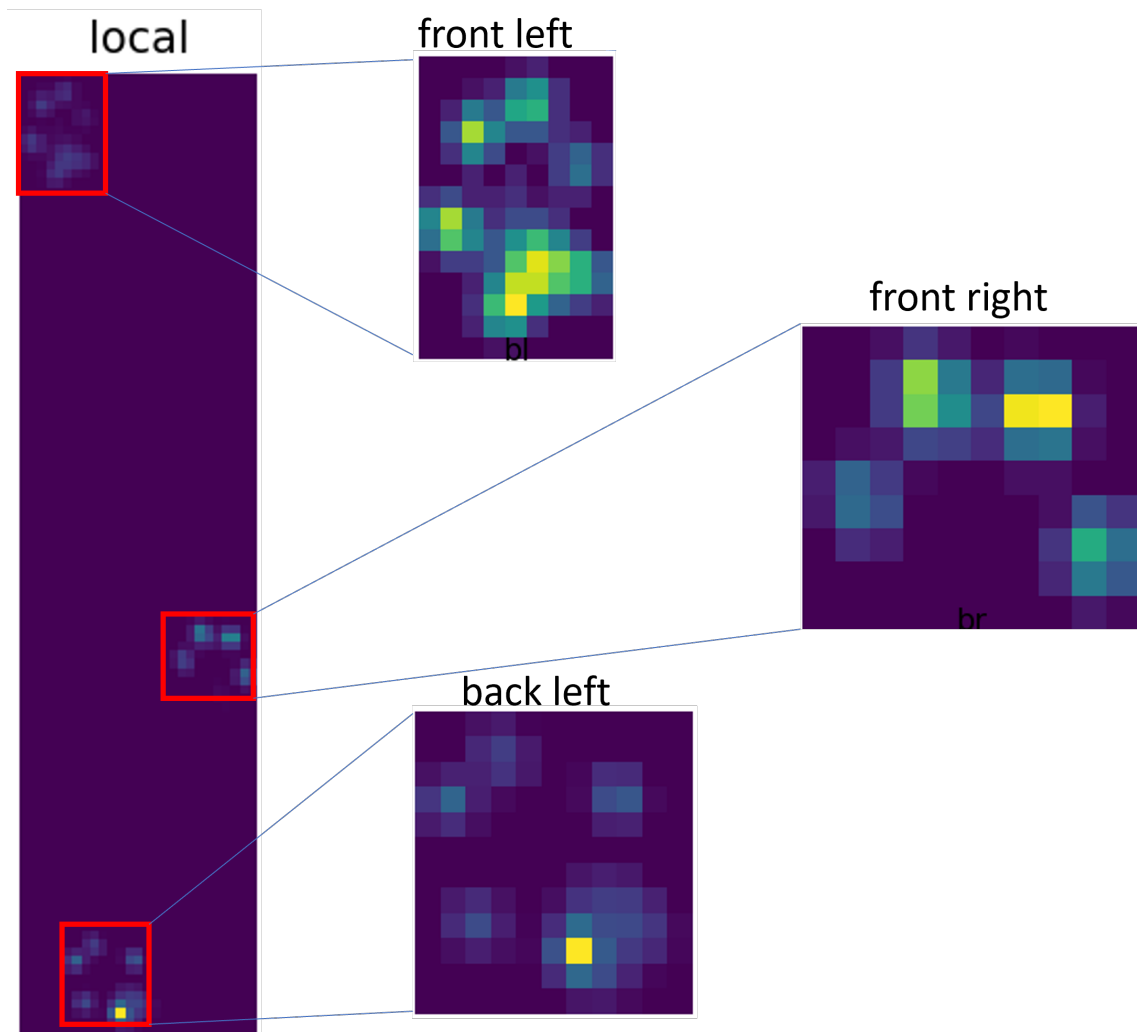


Figure 4.2: Local view with added individual paw enlargements

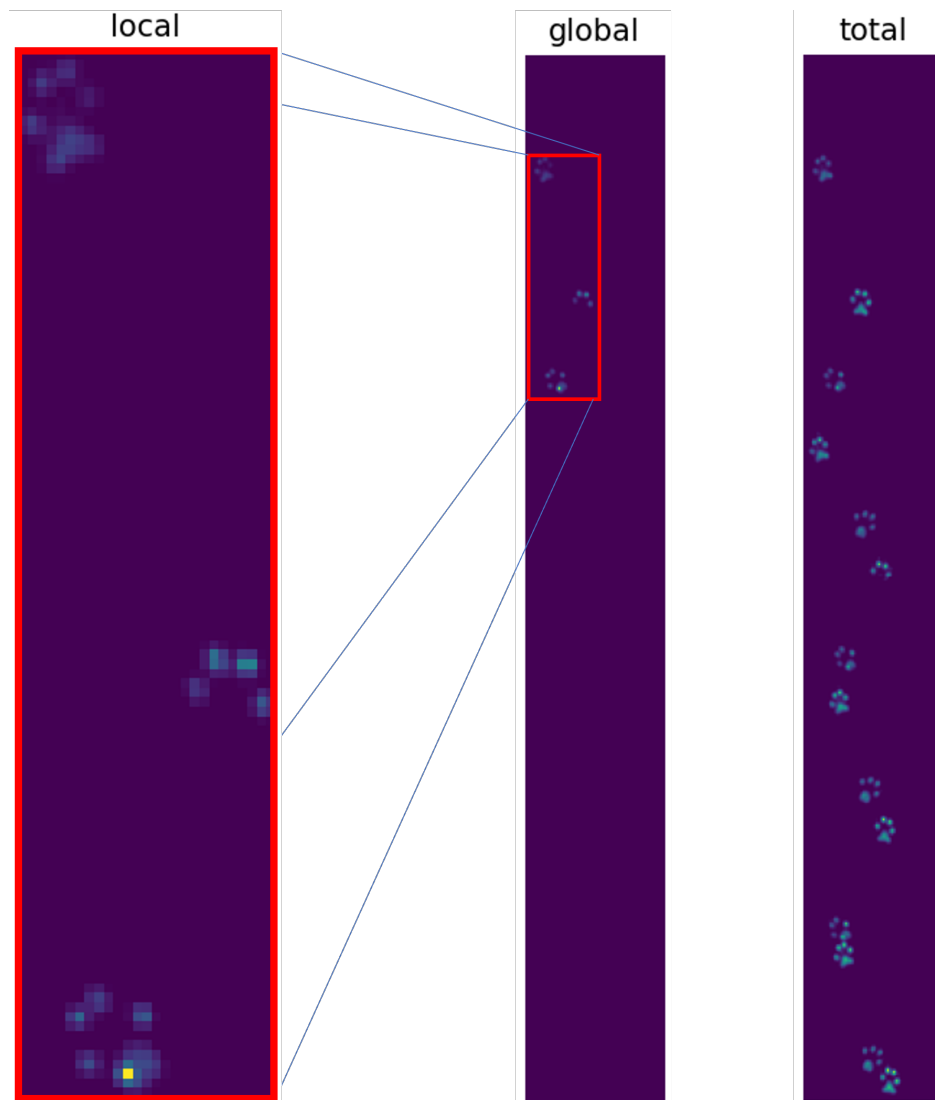


Figure 4.3: Local view embedded in global view. Total view as a layered combination of multiple global views over time

4.2.2 Paw Detection

The main goal of the paw detection is to determine for each walk which paw currently touches the mat, and therefore is represented by the pressure data. This precise distinction is essential for the feature creation, as most features are premised on the progression and behavior of each paw individually during the course of measuring. Confusion or mixing up of the paws has direct and adverse effects on the computation and extraction of the features. For example, an incorrectly detected paw leads to data suggesting that the tested dog either avoids using the (wrongfully) not detected paw or utilizes a constrained walking sequence, which both could potentially be distinct indicators for a musculoskeletal injury. Thus, three approaches for the paw detection were tested, with the combined approach proving most advantageous and resilient to errors.

The assumption for all approaches is that each dog which is measured, loads all its four legs. This means that during every gait cycle, each paw is obliged to have touched the mat, no matter how slight the ground contact might have been. This way, heavily limping dogs and their data still can be analyzed as long as no paw is kept airborne throughout the course of one gait cycle.

All paw detection approaches are based on the same data input for the raw matrix data. With the first paw detection approach, the strategy was to identify the paws at any given moment without consideration of preceding paw detection results or estimations. The mechanics were based solely on the relative position of the paws to each other and the presumed sequence, which served as an indicator which paw is most likely to be lowered next. The idea of using relative paw placement positions originated from the fact that the data set consists of a slower measurement ('walk') and a faster one ('trot') which technically are defined gait types in horse dressage or dog agility [56, 57] but were not strictly followed during data gathering. Without that regard of gait definition of walk and trot, the relative position detection showed a higher error rate and produced inconsistent results for the paw detection.

Therefore, in combination with the above-mentioned assumption, a simplified approach was implemented where the relative position of paws to each other were omitted, and only the paw sequence over time is considered. To further enhance stability, the paw steps were alphabetically encoded, from A through D, by their occurrence in the sequence, independently of their actual function in the movement apparatus. Hence, no additional information about the paws is known, except that paw labeled B always follows step after paw labeled A was set, C follows B et cetera until A follows D and the cycle is repeated. The problem with that approach is that the order of the labels does not always match the actual order of the paws. For instance, labels C and D follow after B, but when the dog places its next two paws simultaneously, it is not clear, which of these two paws was in an earlier gait cycle assigned which label. This problem becomes even

more evident when a previous paw labeled with D contacts the mat earlier than a paw previously labeled C. With that approach, small variances in time of placement can have a significant impact on the paw detection quality.

Moreover, different walks measured also do not always start off the measurement with the same paw. As a result, each paw detection begins the labeling with label A, but across multiple dogs, label A does not correspond to the same paw, preventing meaningful comparison.

The third paw detection approach employed a distinct methodology for identifying the paws. In this approach, the paws are simply designated as either front or back paw. Subsequently, the extracted features are calculated as an average value for both paws in each pair of limbs. To determine whether a paw is front or back, both the time and position of each paw step are compared to preceding and subsequent paw steps. The underlying assumption is that hind paws do not extend beyond front paws during relatively slow gaits such as trot or walk [57]. Consequently, only a front paw can be overtaken by another front paw, but itself overtakes one or both hind paws during a step cycle. All paw steps that, in comparison to paws being planted directly before or after, overtake more paws than they are overtaken are then labelled as front paws. In contrast, the reverse is true for hind paws. Following manual validation, this approach was found to be more reliable than the other strategies in the majority of cases. Measurements in which the difference in step quantity between front and back pairs exceeded two were discarded.

A comparison of the different approaches in terms of their accuracy revealed that the third approach was the most accurate. Since the data were not labeled during acquisition, the verification of the paw detection was done manually with the visualization of the data. Following the paw detection throughout the entire walk and the assignment of paw labels to the data, the next step is the extraction of potentially relevant features that could be used as indicators for irregular or unhealthy walking patterns.

4.2.3 Features

Besides the paw detection which attributes the pressure data to one of the pair of paws, for each dog some additional properties are collected during the measurement, of which some are also used in the further analysis as features themselves. These include weight, height, length, age, sex, and neutering status. A categorized display of the total extracted features and their respective unit or scope can be seen in [Table 4.1](#). In the following, a concise explanation of each feature and its calculation will be given. Beforehand, it is important to mention that each feature is calculated individually per pair of paws, forming an average value for that pair based on the features calculated from the measurements throughout the assessment. For every feature which is calculated by averaging multiple

Spatial	Time	Pressure	Physical attributes
step length [m]	ground contact duration [s]	average pressure [N/cm^2]	age [y]
area of contact [cm^2]			weight [kg]
pace [km/h]			neutered [0/1]
	peak slope descend [N/cm^2s]		sex [0/1]
	peak slope ascend [N/cm^2s]		withers height [cm]
	impulse [Ns]		back length [cm]

Table 4.1: Categorization of features by dimension

values from multiple steps, the standard deviation is also determined and used as an individual feature.

Spatial

Step length is the distance, measured in meters [m], between two consecutive steps of a pair of paws. The spatial distance between two steps of a pair of paws is calculated based on their respective global positions, given by the pressure sensor cell positions, and then converted using a conversion factor that transforms cell size to centimeters. The area of contact, expressed in square centimeters [cm^2], represents the average surface area occupied by a paw from its initial placement to its lifting. This measurement reflects the extent of placement distribution and pressure exertion during each step. Additionally, pace, measured in kilometers per hour [km/h], is the speed of movement of the dog. It is calculated by dividing the total distance covered, usually spanning the entire measurement mat, by the duration of the measurement.

Time

Ground contact duration, also measured in seconds [s], reflects the stance phase of the movement. Additional measures taken are the peak slope ascend and peak slope descend, denoted in Newtons per square centimeter per second [N/cm^2s]. These metrics describe the rate of pressure change from the moment a paw first touches the mat until it reaches its maximum force (ascend) and from the peak until the paw lifts off (descend). Both metrics are calculated by dividing the peak pressure value by the duration spent on the ground before or after the peak, revealing the pressure dynamics of the ascent and descent phases. Another value which was measured is the impulse (in Newton-seconds [Ns]) the paws showed during measurement. The impulse is calculated by multiplying the exerted force on the mat over the duration of the application of that force.

Pressure

The peak pressure, measured in Newtons per square centimeter [N/cm^2], represents the maximal pressure exerted by the paw during a step. This metric serves as a key indicator of weight distribution and limb loading dynamics and is typically observed around the midpoint of the contact time in healthy canines [38]. In order to provide insights into the paw's overall loading behavior, the average pressure, expressed in Newton per square centimeter [N/cm^2], offers an overview of the pressure distribution throughout the entire step. Additionally, the vertical force in Newton [N] is also determined. It is calculated by multiplying the pressure by the area.

Physical attributes

In this section, data distribution across the dogs is stated by the mean average and the standard deviation, as well as the range from minimum to maximum value. The data set consists of dogs averaging 5.93 ± 3.61 years ranging from 1 to a maximum of 14 years, with a median of 5 years. Of the 785 data points, 538 (68.5%) are diagnosed as ill. The weight ranges from 2.2 to 71 *kg* with an average of 19.18 ± 13.05 *kg*. On average, the dogs have a back length of 49.20 ± 11.88 *cm* ranging from 18 to 74 *cm*. The average withers height has an average value of 47.34 ± 13.80 *cm* spanning from 23 to 71 *cm*.

5 Evaluation

This chapter will evaluate the proposed approach both from a performance and a quality perspective.

5.1 Experiments

In this section, the used algorithms and experimental methodology is explained in depth. The data basis of the conducted experiments was the created data set described in [Section 4.2](#).

5.1.1 Preprocessing

In this step, the examination data of the dogs provided by TiHo was prepared for processing. Categorical values, such as medical diagnosis (ill or healthy) and neutering status, were converted to a binary scale without consideration of the specific kind of diagnosis or illness. In order to ensure that the most significant variables were included in the analysis, the relevant features were extracted. Missing values within the dataset were addressed using a simple imputation strategy. This involved replacing each missing value with the median value of the corresponding feature, thus maintaining data integrity and continuity. Random features were added to the data set, which are supposed to be completely irrelevant to the target variable. In this way, it can be determined how well the GA using the classifiers can detect irrelevant features and ignore them for the minimal feature subset. All features were standardized using scikit-learn's 'StandardScaler' function, which subtracts the mean and scales to unit variance. The rationale for this approach is that classifiers such as k-Nearest Neighbors, which employ distance measures for prediction, require standardized data input [17].

5.1.2 Algorithms

Three machine learning classification algorithms were chosen for the experiments: Logistic Regression (LR), k-Nearest Neighbors (KNN), and Support Vector Machine (SVC). The LR classifier models the probability of a binary outcome based on one or more predictor variables. It fits a logistic function to the data, which predicts the probability of the target variable belonging to a particular class. KNN classifiers predict unseen data based on a majority vote of its k nearest neighboring data samples learned from the training set. SVC tries to find the optimal hyperplane to separate different classes in the feature space. If data are not linearly separable in the original feature space, it achieves this by transforming the input data into a higher-dimensional space using kernel functions, enabling the separation of classes. For each of the classifiers, the hyperparameters were tuned as described in 5.1.3.

The 'sklearn-genetic-opt' library was employed to integrate the feature selection process with the genetic algorithm. This library is specifically designed to utilize GAs for optimization and feature selection, making it effective in searching complex solution spaces to identify (near) optimal subsets of features. The integration of feature selection within the GA framework ensures that the chosen subsets enhance the classifier's predictive performance while reducing the number of features. Thereby, addressing the multi-objective problem of maximizing model effectiveness while minimizing feature count. The library allows for the configuration of parameters and probabilities for genetic operators, enabling the fine-tuning of the algorithm to balance exploration and exploitation. The incorporation of fundamental genetic operators, including selection, crossover, and mutation, ensures a robust evolutionary process, which is crucial for the discovery of optimal solutions. Furthermore, the seamless integration with the 'scikit-learn' ecosystem facilitates the utilization of a wide range of classifiers. The compatibility with scikit-learn pipelines further enhances this by allowing the incorporation of preprocessing, feature selection, and model training into a single workflow. A major advantage of 'sklearn-genetic-opt' is its support for cross-validation, which is especially valuable for small datasets like the one at hand. The use of cross-validation with the F1 score as the performance metric ensures a more reliable assessment of model performance, avoiding the pitfalls of a limited data split.

5.1.3 Hyperparameter tuning

The procedures described are identical for all three classifiers, with the sole exception of the possible hyperparameters and their respective ranges of values. The dataset was divided into three distinct subsets, i.e., a training set, a validation set, and a test set. The

hyperparameters of each classifier were optimized using stratified five-fold cross-validation (CV) of the validation set to ensure an equal distribution of classes across both sets. The average F1 score of five runs was used as a metric to identify the optimal parameters identified by 'RandomizedSearchCV' with 10^4 iterations. Random search was selected due to limited computational resources and its higher efficiency and superiority compared to grid search as suggested in [58]. The hyperparameter search spaces for each classifier are detailed in Table 5.1. A log-uniform distribution was used for the C penalty of LR and SVC. That is, to ensure that a wide range of values was considered, while maintaining a focus on smaller values, since metric values are likely to differ more between smaller values than between higher values [59].

When tuning hyperparameters for machine learning algorithms, certain combinations may not be feasible due to inherent constraints of the models. For example, in an SVC, using a polynomial kernel requires specifying the degree of the polynomial, but this hyperparameter is irrelevant and should not be set if a linear kernel is chosen. Similarly, for an SVC, the kernel coefficient 'gamma' is critical for RBF, polynomial, and sigmoid kernels, but it has no effect when a linear kernel is used. In the LR classifier, combining an 'L2' penalty with the 'liblinear' solver is a valid approach, but using an 'elasticnet' penalty is only compatible with the 'saga' solver. During the hyperparameter tuning process, all hyperparameter combinations deemed infeasible were omitted from the evaluation process. The determined optimal hyperparameters can be found in Table 5.1 in the 'Best' column. The previously mentioned parameters are referenced in this thesis when discussing tuned classifiers.

5.1.4 Feature Selection

In order to select the most relevant features for the classifier, a GA was employed. The objective was to reduce the size of the feature subset and to simultaneously maximize the F1 score of the classifier. The initial population was generated randomly to ensure an unbiased starting point. A tournament selection process with five chromosomes was employed to select individuals for reproduction while balancing exploration and exploitation. In order to ensure that the most optimal solutions were retained, an elitism strategy was implemented. The genetic operators included uniform crossover with a probability of $p = 0.7$ to mix parent solutions more evenly and uniform bit flip mutation with a probability of $p = 1/n$, where n is the total number of features, which for this data set was equal to 51. The probability based on the number of features was found to be a suitable fit in numerous problems [14]. Moreover, the population size was 200, and the number of generations was 50.

The GA is employed as a wrapper method to identify the minimal sufficient feature

Algorithm	Hyperparameter	Type	Lower	Upper	Transform	Choice	Best	Default
LR								
	solver	discrete	-	-	-	'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'	'saga'	'lbfgs'
	L1_ratio	numeric	0	1	-	- 'none', 'L1', 'L2', 'elasticnet'	0.93	'none'
	penalty	discrete	-	-	-	'L1', 'L2', 'elasticnet'	'elasticnet'	'L2'
	C value	numeric	-4	4	10^x	-	2.282	1.0
KNN								
	metric	discrete	-	-	-	'euclidean', 'manhattan', 'minkowski'	'manhattan'	'minkowski'
	n_neighbors	integer	1	100	-	-	4	5
	weights	discrete	-	-	-	'uniform', 'distance'	'distance'	'uniform'
	p	discrete	-	-	-	[1,2]	1.623	2
SVC								
	kernel	discrete	-	-	-	'linear', 'poly', 'rbf', 'sigmoid'	'rbf'	'rbf'
	C value	numeric	-4	4	10^x	-	356.713	1.0
	gamma	numeric	-4	4	10^x	-	0.186	'scale'
	degree	integer	1	10	-	-	4	3

Table 5.1: Hyperparameter search spaces of the classifiers. The optimal results are presented after applying the transform function, if applicable. Default parameters are given as of scikit-learn version 1.4.2.

subset while simultaneously optimizing the classifier’s performance. Each chromosome in the GA is of the same length as the total feature subset, with each bit encoding a feature occurrence in binary. A value of 0 indicates that the feature in question is not considered in this particular solution instance, whereas a value of 1 indicates that the feature is included in the subset. Subsequently, the binary string of an individual is translated to a set of features, which utilizes only these features to retrain the classifier. As the fitness, the mean F1 score of stratified 5-fold cross-validation on the training set is employed. Upon reaching the termination criteria, which is the maximum number of generations reached, the feature subset of the overall fittest solution is employed to identify the most relevant features selected by the classifier. Three classifiers were employed, each with two variations: with default and tuned hyperparameters. Each variation of the classifiers was run 31 times within the GA to provide statistically more significant results.

5.2 Results

The results of the experiments are presented in this section. Firstly, the results regarding the predictive performance of three different classifiers, namely Logistic Regression, k-Nearest Neighbor, and Support Vector Classifier, are presented and examined.

Secondly, this section covers the results regarding the research question what conclusions about the orthopedic health of canines can be drawn from the extracted features and specifically which features are considered to be most relevant to the prediction process.

Figure 5.1 shows the average fitness evolution of the GA using the default and tuned classifiers with feature selection. As a baseline comparison, the constant fitness of a naive classifier (NC) is plotted as well. The NC always predicts the majority class of the training sample. As the initial subsets chosen by the GA are generated randomly, the fitness of every classifier variation of the first generation is greater than compared to that of the naive classifier. Of the three proposed classifiers, the LR classifier performs with the least fitness regardless of which hyperparameters, default or tuned, are used (F1-score approximately 0.88). The evolutions of the fitness are almost identical for both variations of the LR classifier, as the graphs are mostly overlapping, with a slightly superior fitness in the default variant. Both LR classifier fitness plots show the lowest fitness increase after GA termination, with improvements of 5.83% and 5.96%, respectively. On the contrary, the tuned variations of the KNN and SVC classifiers show major improvements compared to their default counterparts. Looking at the absolute fitness, both KNN and SVC reach an F1-score of about 0.98. While the tuned KNN classifier increases the fitness on average

by about 11.7% to an F1-score of 0.98, the default version does so by about 12.2%, albeit while reaching a lesser absolute fitness (F1: 0.92). The highest relative increase in fitness and absolute scoring is achieved by the tuned SVC (16.11%), scoring on average 0.982 as the final fitness score. With regard to the research question of identifying which classifier is optimal in predicting ill dogs based on their gait data, it can be said that both KNN and SVC show comparable results and tend to generalize the training data sufficiently to correctly predict on new and unseen data. The LR classifier appears to have difficulties generalizing the data, yet it still has a higher accuracy than the NC.

The findings derived from the fitness evolution are also in line with [Table 5.2](#) in which the mean final performance evaluations on the test set for each classifier can be seen. Both the accuracy and the F1-score are listed for each classifier using either the default or the optimized hyperparameters. Since the NC does not have any parameters, the scores for both default and tuned variation are identical. The performance of the NC is used as a baseline comparison as well. The metrics are based on the performance of the overall fittest chromosome of an individual run of the GA, predicting the target variable on the test set. When using only default parameters, all three classifiers, especially the SVC, show F1-score performances comparable to that of the NC. Both the default LR and KNN are performing best out of the default classifiers, with an F1-score of approximately 0.88 and an accuracy of 0.821 and 0.833, respectively. This changes when using tuned hyperparameters, as the LR classifier has a significantly lower predictive performance compared to KNN and SVC for both of the metrics. The LR classifier stagnates at about the same level regardless of which parameters are set. Both KNN and SVC show a significant increase when tuned, reaching F1-scores of about 0.975 and 0.968, respectively, and similar accuracy scores. This shows the generalization capabilities on unseen data and the fact that both classifiers are as capable of identifying an ill dog, albeit that the KNN is slightly superior.

These evaluation metrics taken from the prediction performance on the test set shows that KNN and SVC are most optimal alike out of the proposed classifiers in detecting unsound canines. Furthermore, it can be seen that, except for the LR classifier, the hyperparameter optimization can have a great impact on the classifier performance. This is most evident for the SVC, as both variations start off in [Figure 5.1](#) at about the same fitness score on average, but the tuned SVC shows much greater improvement over multiple generations. While the LR classifier improves over the run of the GA, it still has a lower performance than the untuned versions of the other classifiers. However, it should be noted that all proposed classifiers are better than the NC comparison baseline, showing that the data set provides data to differentiate between dogs with and without a medical condition.

In order to address which conclusions can be drawn about the orthopedic health of ca-

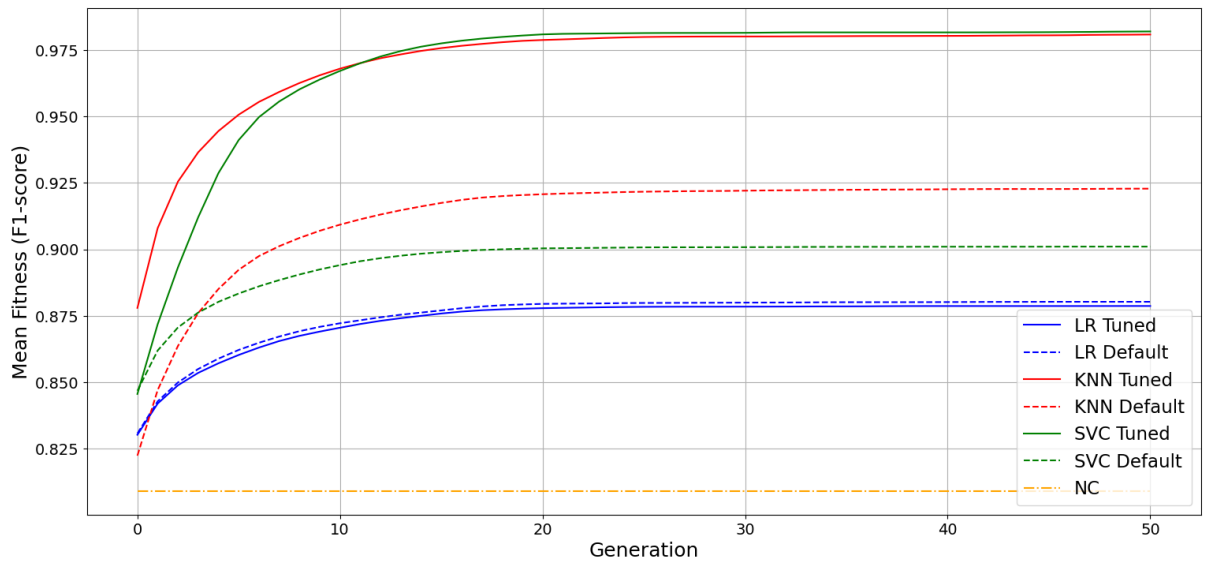


Figure 5.1: Mean fitness evolution of the GA for default and tuned classifiers with feature selection. The performance of the naive classifier plotted as baseline.

Classifier	Default		Tuned	
	Acc	F1	Acc	F1
NC	0.679	0.809	0.679	0.809
LR	0.821	0.876	0.812	0.87
KNN	0.833	0.88	0.967	0.975
SVC	0.746	0.832	0.956	0.968

Table 5.2: Mean performances on the test set of default and tuned classifiers when using the best found feature subset of a single run of the GA.

nines from the extracted features, and specifically which features are considered to be most relevant, [Table 5.3](#) is referred to. The table shows the relative occurrences of features per classifier and the average occurrences across all three classifiers. Features which were not part of the best found subset of a particular classifier at least once are not included in the list for that specific classifier. Random features are highlighted, as they are used as a quality measure of the feature subset selection process. The order of the features represents the frequency of occurrence for a feature selected by the tuned classifiers. Consequently, the higher a feature is listed, the more frequently that feature was part of the final best-found feature subset of the tuned classifier. Moreover, for each classifier, the features and their respective relative occurrences are listed as chosen by the default variation. The used abbreviation suffixes *_F* and *_B* characterize whether that particular feature describe the front (*_F*) or back (*_B*) paws, respectively. The *std_* prefix denominates the standard deviation of that particular feature. A comprehensive overview of the abbreviations used in the feature names can be found in the associated caption of the table.

When analyzing the list of chosen features individually per classifier, it becomes apparent that the LR classifier chose all ten random features at least twice. The most frequently chosen random feature was selected in 58.06% of the runs. Overall, the random features are ranked across a much broader range of the list compared to the other two classifiers. As the random features were included as a quality and reliability measure, at least all features ranking lower than the best ranking random feature (*random_8*) are ignored in the further analysis. The features ranking higher seem to some extent to be coherent with the findings of the other classifiers and related works mentioned in [Chapter 3](#). For example, with *BCS*, *sex*, *weight*, and *age*, four of the features from the 'Physical attributes' category seen in [Table 4.1](#) are present. This is also in line with the other classifiers, as these features are the most frequently chosen ones on average. Moreover, for the LR classifier, both *paw area* and *impulses* features are amongst the top ranking features. These variables are found to be effective in identifying lame limbs, as stated by [28, 42]. On the contrary, both *vertical force* features rank lower than two random features, although the above studies consider these to be reliable as well.

The tuned KNN classifier chose four random features at least once, but it is noteworthy that three out of these were chosen only a single time. Meaning that the KNN in few cases might be considering unnecessary noise to predict the target variable, but does so rarely. In contrast, the default KNN classifier selected only two of these random features at least once while still having a lower predictive performance, as demonstrated in [Figure 5.1](#) and [Table 5.2](#). When evaluating the most often selected features for the tuned KNN classifier, most of the features from the 'Physical attributes' category are ranked as the most relevant features as well. While the KNN classifier assigns relatively high rankings

LR	Tuned	Default	KNN	Tuned	Default	SVC	Tuned	Default	Mean	Tuned	Default
Feature	% Occ.	% Occ.	Feature	% Occ.	% Occ.	Feature	% Occ.	% Occ.	Feature	% Occ.	% Occ.
BCS	100.00	100.00	age	100.00	93.55	age	100.00	100.00	BCS	100.00	100.00
sex	100.00	100.00	BCS	100.00	100.00	BCS	100.00	100.00	weight	100.00	97.85
weight	100.00	100.00	length	100.00	51.61	length	100.00	100.00	sex	97.85	86.02
des_peak_slopes_B	93.55	80.65	weight	100.00	93.55	paw_area_F	100.00	74.19	age	90.32	89.25
pace	93.55	96.77	sex	93.55	58.06	sex	100.00	100.00	paw_area_F	86.02	78.49
paw_area_F	83.87	100.00	paw_area_F	74.19	61.29	weight	100.00	100.00	length	82.80	68.82
std_impulses_F	83.87	87.10	avg_pres_F	67.74	93.55	neutered	96.77	83.87	impulses_F	74.19	56.99
impulses_F	80.65	80.65	impulses_F	61.29	51.61	std_avg_pres_B	93.55	16.13	paw_area_B	69.89	51.61
step_lengths_B	80.65	80.65	impulses_B	51.61	51.61	paw_area_B	90.32	45.16	std_impulses_F	60.22	58.06
impulses_B	77.42	87.10	avg_pres_B	48.39	67.74	std_contact_dur_F	90.32	74.19	impulses_B	58.06	55.91
age	70.97	74.19	paw_area_B	48.39	41.94	impulses_F	80.65	38.71	std_contact_dur_F	48.39	51.61
paw_area_B	70.97	67.74	std_impulses_F	48.39	58.06	vert_forces_F	67.74	32.26	vert_forces_F	47.31	46.24
std_step_lengths_B	67.74	61.29	std_des_peak_slopes_B	45.16	19.35	des_peak_slopes_F	48.39	22.58	std_avg_pres_B	43.01	18.28
std_asc_peak_slopes_B	61.29	58.06	std_asc_peak_slopes_B	38.71	48.39	std_impulses_F	48.39	29.03	step_lengths_B	41.94	77.42
random_8	58.06	74.19	std_step_lengths_B	38.71	61.29	impulses_B	45.16	29.03	neutered	40.86	37.63
std_paw_area_B	58.06	64.52	std_avg_pres_B	35.48	35.48	vert_forces_B	41.94	35.48	des_peak_slopes_B	39.78	56.99
std_paw_area_F	54.84	74.19	std_contact_dur_F	35.48	51.61	avg_pres_B	35.48	35.48	std_step_lengths_B	39.78	74.19
std_vert_forces_F	51.61	70.97	std_impulses_B	35.48	45.16	std_impulses_B	35.48	19.35	std_asc_peak_slopes_B	37.63	44.09
contact_dur_F	48.39	61.29	vert_forces_F	35.48	77.42	std_avg_pres_F	25.81	70.97	std_impulses_B	36.56	31.18
length	48.39	54.84	step_lengths_F	32.26	54.84	std_vert_forces_B	22.58	19.35	avg_pres_B	35.48	55.91
step_lengths_F	48.39	67.74	vert_forces_B	29.03	51.61	std_des_peak_slopes_B	19.35	9.68	avg_pres_F	35.48	49.46
random_1	41.94	48.39	des_peak_slopes_B	25.81	48.39	step_lengths_B	19.35	90.32	vert_forces_B	34.41	38.71
std_contact_dur_B	41.94	38.71	neutered	25.81	12.90	std_asc_peak_slopes_F	16.13	16.13	step_lengths_F	32.26	59.14
std_des_peak_slopes_F	38.71	70.97	std_des_peak_slopes_F	25.81	25.81	step_lengths_F	16.13	54.84	pace	31.18	56.99
std_impulses_B	38.71	29.03	step_lengths_B	25.81	61.29	asc_peak_slopes_B	12.90	54.84	std_vert_forces_B	27.96	46.24
std_vert_forces_B	38.71	64.52	std_vert_forces_B	22.58	54.84	asc_peak_slopes_F	12.90	54.84	des_peak_slopes_F	26.88	17.20
vert_forces_F	38.71	29.03	asc_peak_slopes_B	16.13	58.06	avg_pres_F	12.90	35.48	std_paw_area_B	23.66	33.33
vert_forces_B	32.26	29.03	std_asc_peak_slopes_F	16.13	19.35	std_asc_peak_slopes_B	12.90	25.81	std_des_peak_slopes_B	22.58	15.05
asc_peak_slopes_B	29.03	9.68	asc_peak_slopes_F	12.90	35.48	std_step_lengths_B	12.90	100.00	contact_dur_F	21.51	78.49
random_7	29.03	22.58	contact_dur_F	12.90	74.19	std_vert_forces_F	6.45	48.39	std_des_peak_slopes_F	21.51	41.94
avg_pres_F	25.81	19.35	std_contact_dur_B	12.90	51.61	contact_dur_B	3.23	12.90	std_vert_forces_F	21.51	55.91
des_peak_slopes_F	25.81	16.13	contact_dur_B	9.68	29.03	contact_dur_F	3.23	100.00	random_8	20.43	26.88
std_step_lengths_F	25.81	16.13	std_paw_area_B	9.68	12.90	std_paw_area_B	3.23	22.58	asc_peak_slopes_B	19.35	40.86
asc_peak_slopes_F	22.58	9.68	des_peak_slopes_F	6.45	12.90	des_peak_slopes_B	0.00	41.94	std_avg_pres_F	18.28	40.86
avg_pres_B	22.58	64.52	random_4	6.45	0.00	height	0.00	19.35	std_contact_dur_B	18.28	50.54
random_5	22.58	9.68	std_avg_pres_F	6.45	25.81	pace	0.00	41.94	std_paw_area_F	18.28	59.14
random_6	22.58	9.68	std_vert_forces_F	6.45	48.39	random_1	0.00	3.23	std_asc_peak_slopes_F	17.20	20.43
std_avg_pres_F	22.58	25.81	random_5	3.23	3.23	random_2	0.00	3.23	asc_peak_slopes_F	16.13	33.33
random_4	19.35	9.68	random_6	3.23	0.00	random_4	0.00	3.23	random_1	13.98	17.20
std_asc_peak_slopes_F	19.35	25.81	random_8	3.23	3.23	random_5	0.00	3.23	random_7	9.68	7.53
std_contact_dur_F	19.35	29.03	height	0.00	32.26	random_8	0.00	3.23	random_4	8.60	4.30
random_9	12.90	29.03	pace	0.00	32.26	random_9	0.00	22.58	random_5	8.60	5.38
random_10	9.68	16.13	std_paw_area_F	0.00	32.26	std_contact_dur_B	0.00	61.29	random_6	8.60	3.23
contact_dur_B	6.45	19.35	std_step_lengths_F	0.00	12.90	std_des_peak_slopes_F	0.00	29.03	std_step_lengths_F	8.60	34.41
random_2	6.45	22.58				std_paw_area_F	0.00	70.97	contact_dur_B	6.45	20.43
random_3	6.45	6.45				std_step_lengths_F	0.00	74.19	random_9	4.30	17.20
std_des_peak_slopes_B	3.23	16.13							random_10	3.23	5.38
height	0.00	32.26							random_2	2.15	8.60
neutered	0.00	16.13							random_3	2.15	2.15
std_avg_pres_B	0.00	3.23							height	0.00	27.96

Table 5.3: Relative feature occurrences of default and tuned classifiers; grouped by classifier and sorted by occurrence of feature. asc:ascend, avg:average, _B:back paws, BCS:body condition score, des:descend, dur:duration, _F:front paws, pres:pressure, std:standard deviation, vert:vertical

to the features *paw area*, *impulses*, and *average pressure*, these are selected significantly less frequently (ranging from 48.39-74.19%) than the previously mentioned highest-ranked features. It is somewhat surprising that the highest occurrence of a non-physical attribute is only at 74.19% (*front paw area*), as the machine learning algorithm has no possibility of determining whether a dog's gait is skewed if no gait feature is taken into consideration. If the classifier were to employ only physical attributes for prediction, then the algorithm would generalize the underlying training data to recognize body shapes and dimensions of ill dogs, rather than their specific gait attributes.

The tuned SVC is the only classifier that did not select a random feature once. Moreover, it has the lowest total number of distinct features selected (33) and the highest number of features selected in each iteration of the GA (6). These include *age*, *BCS*, *body length*, *front paw area*, *sex*, and *weight*. With that, five out of six features from the 'Physical attributes' category seen in [Table 4.1](#) are present. It should be noted that *height* is not once part of the feature subset for any of the tuned classifiers, meaning that it can be considered irrelevant with regard to the target prediction. Of the 33 total features the tuned SVC chose at least once, 10 have a frequency of more than 90%. In addition to the physical attributes of the canine, two features are derived from the gait data for each pair of limbs. Only the *paw area* ranks that high for both pair of limbs, emphasizing the importance of that feature. The features for the standard deviation of the *back limb average pressure* and *front contact duration* are the only standard deviation-features amongst the features which are chosen in more than 90% of the runs. Especially the tuned KNN shows no high ranking of the standard deviation features, as the highest relative occurrence is 48.39% for the *front impulses*.

In total, only *BCS* and *weight* are selected in every feature subset. This is plausible given that both are chosen as they are closely related and influence each other. Across all three classifiers, the majority of the physical attributes are ranking highest. This could be interpreted as a form of normalization of the gait data between different sized canines as the physical attributes give insight into the normalized body dimensions when accounting for factors like weight, size, or length as suggested by some studies [33, 36, 37]. However, every classifier employed, regardless of the hyperparameters utilized, showed improvements in fitness over multiple generations of feature selection. When optimizing the hyperparameters, the maximum fitness and therefore quality of solution produced superior results, except for the LR classifier. The final performance evaluation showed no significant difference in F1-scores or accuracy between KNN and SVC, implying their equality in predicting ill canines.

5.3 Limitations

This thesis provides insights into the use of genetic algorithms for feature selection in gait analysis. However, it is important to acknowledge some limitations. Firstly, a limited set of three classifiers is employed. These classifiers include Logistic Regression, k-Nearest Neighbor, and Support Vector Classifier. Hence, it can be suggested that a broader range of classifiers could be evaluated to potentially improve the accuracy and robustness of the results and provide a more diverse insight into classifier prediction performances. Secondly, the analysis combines data from both paws at the front or at the back, respectively, which might obscure differences specific to each paw individually. Thirdly, the dataset itself is limited in size, which could affect the generalizability of the findings, especially for unseen breeds, diseases, or gait patterns.

Furthermore, this thesis does not differentiate between various illnesses, treating all instances homogeneously, which may result in the overlooking of specific patterns associated with different conditions. The absence of inclusion of video data and multiple angles might limit the depth of the analysis, as visual information could provide additional insights into gait abnormalities. Finally, the data used in this thesis had to be filtered and preprocessed due to the lack of a clear data acquisition process and standardized guidelines for gait measurements, leading to an even smaller data set. In future research, it would be beneficial to address these limitations in order to enhance the validity and applicability of the findings.

6 Conclusion and Future Work

All three proposed classifiers demonstrated superior performance to a naive classifier implementation, which predicts only the majority class. While the LR classifier showed some improvement after applying feature selection, the tuned hyperparameters could not increase performance significantly. In contrast, both KNN and SVC achieved maximum performance on the test set with accuracy and F1-score values of approximately 0.96 and 0.97, respectively. This suggests that both classifiers are suitable in canine gait analysis with the features extracted from the given data set. For the default variations of the classifiers, KNN showed better performance in predicting the target value and increased its performance with feature selection more. When taking also the selected feature subsets into account, the SVC seems to be superior and more robust, as the tuned version disregarded all random features. Given that the LR classifier selected all random features at least once, the feature occurrences of that classifier have to be interpreted with caution, as some random features were chosen in more than 40% of the runs, suggesting an imprecise or even unreliable feature selection process. The KNN classifier selected some random features rarely, yet the overall predictive performance did not differ greatly from that of the tuned SVC. Contrary to some studies covered in [Subsection 2.1.2](#), the *vertical force* feature was chosen in less than 70% of the runs, revealing a lesser significance than previously expected. While the performances of both tuned KNN and SVC are very much comparable, differences in feature selection are identifiable. Firstly, the KNN classifier selected four random features at least once in the feature subset, while the SVC discarded all of them. This indicates that the SVC is more robust in filtering noisy features from the data, resulting in a more compact feature subset. Secondly, the SVC selected in total more features in every single run of the GA (6) as compared to the KNN classifier (4). When comparing features which are selected at least 90% of the time, the difference is even bigger (KNN: 5, SVC: 10). In other words, the SVC chooses the same features more reliably based on the feature selection than the KNN. The latter has a bigger variance in different features chosen, impeding the selection of the most relevant features in total. With regard to the research question, which of the proposed classifiers exhibits the highest predictive performance, both KNN and SVC are significantly superior compared to the LR classifier, reaching F1-scores of about 0.98 and 0.97, respectively. When taking into consideration the occurrences of selected features in order to generalize the data, the SVC

seems to be more robust against noisy data.

As to the second research question, which features are deemed most relevant for predicting orthopedic diseases, the most selected features in the experiments conducted are predominantly of the 'Physical attributes' category, including weight, BCS, age, sex, and length. Other features selected most frequently, especially by the two best performing classifiers, are paw area, average pressure and their standard deviation as well as the impulse. These features are also consistent with related works on canine gait analysis, as particularly paw area and the impulse are mentioned in multiple studies as reliable features in detecting abnormal gait patterns.

In the following, several possibilities for future research are suggested to enhance the understanding of canine gait analysis through machine learning methods. A crucial first step would be to gather more data from a larger group of dogs. This would help mitigate the limitations of the current limited dataset and enhance the generalizability of the results. Future studies should focus on differentiating between various illnesses or breeds of dogs, with a specific aim to enhance the prediction accuracy for specific orthopedic diseases. By improving the recognition performance associated with particular conditions, it could potentially assist veterinarians in medical examinations. Another promising direction for future work is to conduct experimental studies to validate the algorithm's performance in real-world settings. This could involve applying the best-found features from this thesis to evaluate unseen dogs and assess whether the model can accurately detect illnesses or possibly even the affected limb. Experiments like these would provide critical insights into the practical utility of the proposed methods and their effectiveness in real-world use cases. Additionally, integrating video data and joint angles or joint movement in the analysis could further enhance accuracy of the models, offering a more comprehensive understanding of gait abnormalities. Furthermore, instead of using a binary outcome for lameness detection, future research could develop a more nuanced approach by measuring the severity of lameness on a spectrum. This would provide more detailed information about the extent of gait issues, allowing for a more nuanced examination decision and aid veterinarians in their decision process. Overall, these proposed future work suggestions aim to expand the current applicability and ultimately create more reliable tools for veterinary applications.

Bibliography

- [1] A. Huxley, 1934. <https://www.azquotes.com/quote/140016> [Accessed: 04.06.2024].
- [2] A. Greco, L. Meomartino, G. Gnudi, A. Brunetti, and M. Di Giancamillo, “Imaging techniques in veterinary medicine. Part II: Computed tomography, magnetic resonance imaging, nuclear medicine,” *European Journal of Radiology Open*, vol. 10, p. 100467, Dec. 2022.
- [3] J. C. Lattimer, “Nuclear Medicine Imaging in Animals - Clinical Pathology and Procedures,” Nov. 2019.
- [4] J. C. Lattimer, “Magnetic Resonance Imaging in Animals - Clinical Pathology and Procedures,” Nov. 2019.
- [5] D. Yitbarek and G. G. Dagnaw, “Application of Advanced Imaging Modalities in Veterinary Medicine: A Review,” *Veterinary Medicine : Research and Reports*, vol. 13, pp. 117–130, May 2022.
- [6] B. M. Wiles, A. M. Llewellyn-Zaidi, K. Evans, D. G. O’Neill, and T. W. Lewis, “Large-scale survey to estimate the prevalence of disorders for 192 Kennel Club registered breeds,” *Canine Genetics and Epidemiology*, vol. 4, p. 8, Sept. 2017.
- [7] D. G. O’Neill, H. James, D. C. Brodbelt, D. B. Church, and C. Pegram, “Prevalence of commonly diagnosed disorders in UK dogs under primary veterinary care: results and applications,” *BMC Veterinary Research*, vol. 17, p. 69, Feb. 2021.
- [8] B. P. Hospital, “2016 State of Pet Health Report,” tech. rep., Banfield Pet Hospital, Portland, OR, USA, Apr. 2016.
- [9] C. J. Klinger, *Analyse des Fallaufkommens in deutschen Tierarztpraxen*. PhD thesis, Ludwig-Maximilians-Universitaet Muenchen, Muenchen, 2016.
- [10] M. Melanie, *An Introduction to Genetic Algorithms*. MIT Press, 5 ed., 1999.
- [11] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*. Berlin, Heidelberg: Springer, 3 ed., 1996.

- [12] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, ch. Reproductive Plans and Genetic Operators, pp. 89–120. MIT Press, 1992.
- [13] C. A. C. Coello, G. B. Lamont, and D. A. V. Veldhuizen, *Evolutionary Algorithms for Solving Multi-Objective Problems Second Edition*. Springer Science & Business Media, 2 ed., Sept. 2007.
- [14] R. Kruse, S. Mostaghim, C. Borgelt, C. Braune, and M. Steinbrecher, “Elements of Evolutionary Algorithms,” in *Computational Intelligence: A Methodological Introduction* (R. Kruse, S. Mostaghim, C. Borgelt, C. Braune, and M. Steinbrecher, eds.), pp. 255–285, Cham: Springer International Publishing, 2022.
- [15] A. Deniz, H. E. Kiziloz, T. Dokeroglu, and A. Cosar, “Robust multiobjective evolutionary feature subset selection algorithm for binary classification using machine learning techniques,” *Neurocomputing*, vol. 241, pp. 128–146, June 2017.
- [16] J. Han, J. Pei, and H. Tong, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, July 2022. Google-Books-ID: NR1oEAAAQBAJ.
- [17] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.
- [18] G. John, R. Kohavi, and K. Pfleger, “Irrelevant Features and the Subset Selection Problem,” *In International Conference on Machine Learning*, vol. 129, Feb. 1998.
- [19] R. Kohavi and G. John, “Wrappers for feature selection,” *Artificial Intelligence - AI*, vol. 1, Jan. 1997.
- [20] V. Bolón-Canedo, N. Sánchez-Marroño, and A. Alonso-Betanzos, *Feature Selection for High-Dimensional Data*. Artificial Intelligence: Foundations, Theory, and Algorithms, Cham: Springer International Publishing, 2015.
- [21] A. L. Blum and P. Langley, “Selection of relevant features and examples in machine learning,” *Artificial Intelligence*, vol. 97, pp. 245–271, Dec. 1997.
- [22] D. Millis and D. Levine, *Canine Rehabilitation and Physical Therapy*. Elsevier Health Sciences, Oct. 2013. Google-Books-ID: ObIKAQAAQBAJ.
- [23] E. M. G. Jr and T. R. Brown, *K-9 Structure & Terminology*. Dogwise Publishing, Feb. 2017. Google-Books-ID: L0VADgAAQBAJ.
- [24] R. L. Gillette and T. C. Angle, “Recent developments in canine locomotor analysis: A review,” *The Veterinary Journal*, vol. 178, pp. 165–176, Nov. 2008.

- [25] B. J. Carr and D. L. Dycus, “Canine Gait Analysis,” *Today’s Veterinary Practice*, Feb. 2016.
- [26] E. Surer, A. Cereatti, M. A. Evangelisti, G. Paolini, U. Della Croce, and M. L. Manunta, “A Canine Gait Analysis Protocol for Back Movement Assessment in German Shepherd Dogs,” *Veterinary Sciences*, vol. 7, p. 26, Mar. 2020. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.
- [27] J. M. Vilar, M. Rubio, J. M. Carrillo, A. M. Domínguez, A. Mitat, and M. Batista, “Biomechanic characteristics of gait of four breeds of dogs with different conformations at walk on a treadmill,” *Journal of Applied Animal Research*, vol. 44, pp. 252–257, Jan. 2016. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/09712119.2015.1031778>.
- [28] M. Oosterlinck, T. Bosmans, F. Gasthuys, I. Polis, B. V. Ryssen, J. Dewulf, and F. Pille, “Accuracy of pressure plate kinetic asymmetry indices and their correlation with visual gait assessment scores in lame and nonlame dogs,” *American Journal of Veterinary Research*, vol. 72, pp. 820–825, June 2011. Publisher: American Veterinary Medical Association Section: American Journal of Veterinary Research.
- [29] M. Brønniche Møller Nielsen, T. Pedersen, A. Mouritzen, A. D. Vitger, L. N. Nielsen, H. H. Poulsen, and J. E. Miles, “Kinetic gait analysis in healthy dogs and dogs with osteoarthritis: An evaluation of precision and overlap performance of a pressure-sensitive walkway and the use of symmetry indices,” *PLoS ONE*, vol. 15, p. e0243819, Dec. 2020.
- [30] J. M. Vilar, G. Ramirez, G. Spinella, and A. Martinez, “Kinematic characteristics of myositis ossificans of the semimembranosus muscle in a dog,” *Can Vet J*, vol. 51, pp. 289–292, Mar. 2010.
- [31] C. E. DeCamp, C. M. Riggs, N. B. Olivier, J. G. Hauptman, H. A. Hottinger, and R. W. Soutas-Little, “Kinematic evaluation of gait in dogs with cranial cruciate ligament rupture,” *American Journal of Veterinary Research*, vol. 57, pp. 120–126, Jan. 1996. Publisher: American Veterinary Medical Association Section: American Journal of Veterinary Research.
- [32] D. J. Jevens, J. G. Hauptman, C. E. DeCamp, S. C. Budsberg, and R. W. Soutas-Little, “Contributions to variance in force-plate analysis of gait in dogs,” *American Journal of Veterinary Research*, vol. 54, pp. 612–615, Apr. 1993. Publisher: American Veterinary Medical Association Section: American Journal of Veterinary Research.

- [33] A. Humphries, A. F. Shaheen, and C. B. Gómez Álvarez, “Biomechanical comparison of standing posture and during trot between German shepherd and Labrador retriever dogs,” *PLoS One*, vol. 15, p. e0239832, Oct. 2020.
- [34] K. Voss, L. Galeandro, T. Wiestner, M. Haessig, and P. M. Montavon, “Relationships of Body Weight, Body Size, Subject Velocity, and Vertical Ground Reaction Forces in Trotting Dogs,” *Veterinary Surgery*, vol. 39, no. 7, pp. 863–869, 2010. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1532-950X.2010.00729.x>.
- [35] M. A. Fahie, J. C. Cortez, M. Ledesma, and Y. Su, “Pressure Mat Analysis of Walk and Trot Gait Characteristics in 66 Normal Small, Medium, Large, and Giant Breed Dogs,” *Frontiers in Veterinary Science*, vol. 5, 2018.
- [36] K. Voss, T. Wiestner, L. Galeandro, M. Hässig, and P. Montavon, “Effect of dog breed and body conformation on vertical ground reaction forces, impulses, and stance times,” *Veterinary and comparative orthopaedics and traumatology : V.C.O.T.*, vol. 24, pp. 106–12, Jan. 2011.
- [37] R. M. Alexander and A. S. Jayes, “A dynamic similarity hypothesis for the gaits of quadrupedal mammals,” *Journal of Zoology*, vol. 201, no. 1, pp. 135–152, 1983. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-7998.1983.tb04266.x>.
- [38] H. Miao, Z. Wang, J. Liu, Z. Qian, L. Ren, and L. Ren, “Biomechanical Functions of the Canine Metacarpal and Metatarsal Pads during Locomotion: A Comparative Analysis,” *J Bionic Eng*, vol. 17, pp. 580–590, May 2020.
- [39] W. T. Kano, S. C. Rahal, F. S. Agostinho, L. R. Mesquita, R. R. Santos, F. O. B. Monteiro, M. S. Castilho, and A. Melchert, “Kinetic and temporospatial gait parameters in a heterogeneous group of dogs,” *BMC Veterinary Research*, vol. 12, no. 1, 2016. Publisher: BMC.
- [40] G. J. Jenkins, C. H. Hakim, N. N. Yang, G. Yao, and D. Duan, “Automatic characterization of stride parameters in canines with a single wearable inertial sensor,” *PLOS ONE*, vol. 13, p. e0198893, June 2018. Publisher: Public Library of Science.
- [41] R. L. Bennett, C. E. DeCamp, G. L. Flo, J. G. Hauptman, and M. Stajich, “Kinematic gait analysis in dogs with hip dysplasia,” *American Journal of Veterinary Research*, vol. 57, pp. 966–971, July 1996. Publisher: American Veterinary Medical Association Section: American Journal of Veterinary Research.
- [42] T. LeQuang, P. Maitre, T. Roger, and E. Viguier, “Is a Pressure Walkway System Able to Highlight a Lameness in Dog?,” in *6th World Congress of Biomechanics*

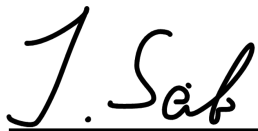
- (*WCB 2010*). *August 1-6, 2010 Singapore* (C. T. Lim and J. C. H. Goh, eds.), IFMBE Proceedings, (Berlin, Heidelberg), pp. 190–193, Springer, 2010.
- [43] R. L. Gillette and T. C. Angle, “Recent developments in canine locomotor analysis: a review,” *Vet J*, vol. 178, pp. 165–176, Nov. 2008.
- [44] C. Ladha, J. O’Sullivan, Z. Belshaw, and L. Asher, “GaitKeeper: A System for Measuring Canine Gait,” *Sensors*, vol. 17, p. 309, Feb. 2017. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.
- [45] S. Jadhav, H. He, and K. Jenkins, “Information gain directed genetic algorithm wrapper feature selection for credit rating,” *Applied Soft Computing*, vol. 69, pp. 541–553, Aug. 2018.
- [46] S. Chatterjee and A. Bhattacharjee, “Genetic algorithms for feature selection of image analysis-based quality monitoring model: An application to an iron mine,” *Engineering Applications of Artificial Intelligence*, vol. 24, pp. 786–795, Aug. 2011.
- [47] L. D. Vignolo, D. H. Milone, and J. Scharcanski, “Feature selection for face recognition based on multi-objective evolutionary wrappers,” *Expert Systems with Applications*, vol. 40, pp. 5077–5084, Oct. 2013.
- [48] L. Vignolo, D. Milone, C. Behaine, and J. Scharcanski, “An evolutionary wrapper for feature selection in face recognition applications,” in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1286–1290, Oct. 2012. ISSN: 1062-922X.
- [49] H. Faris, A. M. Al-Zoubi, A. A. Heidari, I. Aljarah, M. Mafarja, M. A. Hassonah, and H. Fujita, “An intelligent system for spam detection and identification of the most relevant features based on evolutionary Random Weight Networks,” *Information Fusion*, vol. 48, pp. 67–83, Aug. 2019.
- [50] G. N. Karagoz, A. Yazici, T. Dokeroglu, and A. Cosar, “A new framework of multi-objective evolutionary algorithms for feature selection and multi-label classification of video data,” *Int. J. Mach. Learn. & Cyber.*, vol. 12, pp. 53–71, Jan. 2021.
- [51] H.-H. Hsu, C.-W. Hsieh, and M.-D. Lu, “Hybrid feature selection by combining filters and wrappers,” *Expert Systems with Applications*, vol. 38, pp. 8144–8150, July 2011.
- [52] L. Yu and H. Liu, “Feature selection for high-dimensional data: a fast correlation-based filter solution,” in *Proceedings of the Twentieth International Conference on International Conference on Machine Learning, ICML’03*, (Washington, DC, USA), pp. 856–863, AAAI Press, Aug. 2003.

- [53] E. Cantú-Paz, “Feature Subset Selection, Class Separability, and Genetic Algorithms,” in *Genetic and Evolutionary Computation – GECCO 2004* (K. Deb, ed.), (Berlin, Heidelberg), pp. 959–970, Springer, 2004.
- [54] J. Bala, K. D. Jong, J. Huang, H. Vafaie, and H. Wechsler, “Using Learning to Facilitate the Evolution of Features for Recognizing Visual Concepts,” *Evolutionary Computation*, vol. 4, pp. 297–311, Sept. 1996.
- [55] L. Yang and A. Shami, “On hyperparameter optimization of machine learning algorithms: Theory and practice,” *Neurocomputing*, vol. 415, pp. 295–316, Nov. 2020.
- [56] B. J. Carr and D. L. Dycus, “Canine Gait Analysis,” *Today’s Veterinary Practice*, Feb. 2016.
- [57] E. M. G. Jr and T. R. Brown, *K-9 Structure & Terminology*. Dogwise Publishing, Feb. 2017. Google-Books-ID: L0VADgAAQBAJ.
- [58] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.
- [59] P. Probst, B. Bischl, and A.-L. Boulesteix, “Tunability: Importance of Hyperparameters of Machine Learning Algorithms,” Oct. 2018. arXiv:1802.09596 [stat].

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Bachelorarbeit selbstständig und ausschließlich unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Stellen sind als solche kenntlich gemacht.

Die Arbeit wurde bisher in gleicher oder ähnlicher Form weder einer anderen Prüfungsbehörde vorgelegt oder noch anderweitig veröffentlicht.



Unterschrift

11.06.2024

Datum

Statement of Authorship

I herewith assure that I wrote the present thesis independently, that the thesis has not been partially or fully submitted as graded academic work and that I have used no other means than the ones indicated. I have indicated all parts of the work in which sources are used according to their wording or to their meaning.

I am aware of the fact that violations of copyright can lead to injunctive relief and claims for damages of the author as well as a penalty by the law enforcement agency.

J. Sab

Signature

11.06.2024

Date