How molecular diagnostics based on gene expressions can improve diagnostics in Sarcoma

Alma Lennartsson (BME20), Emma Friberg (BME20)

***Abstract*—Sarcoma, a rare and heterogeneous cancer type, present significant diagnostic challenges due to its numerous subtypes. Molecular diagnostics and machine learning models have emerged as promising tools to enhance sarcoma diagnosis. This study, conducted at Qlucore, aims to explore the usage of these new techniques in improving sarcoma diagnostics, with a specific focus on soft tissue sarcomas.**

The primary purpose of the study is to investigate the classification of different subtypes of soft tissue sarcoma based on gene expression analysis. This is performed by investigating and evaluating various classification methods. The study also explores alternative approaches for achieving accurate classification.

The results of this study demonstrate promising potential for the clinical use of molecular diagnostics in accurately diagnosing specific subtypes of sarcoma. However, certain sarcoma subgroups present challenges in classification. The study suggests the adoption of hierarchical classifiers as a potential solution for this. Furthermore, the study emphasizes that the choice of algorithm significantly impacts classification outcomes.

I. Introduction

THE purpose of this study is to investigate how new techniques, such as molecular diagnostics and machine learning models, can be used to improve the diagnostics of sarcoma. The goal is to analyze how well different subtypes of soft tissue sarcoma can be classified based on their gene expression, as well as investigating different ways of doing the classification.

A. Epidemiology of sarcoma

Sarcomas are a rare type of tumour that originates from soft tissue and bone, representing less than 1% of the cancer malignancies worldwide. Sarcomas can be divided into over 100 different subtypes (Table I), and more are constantly discovered [1]. In this study the focus will be on the subtypes of soft tissue sarcomas, which corresponds to 70-80% of all sarcomas. Soft tissue sarcomas can form almost anywhere in the body, but most common are arms, legs, abdomen and retroperitoneum. Soft tissue refers to muscles, tendons, ligaments, cartilage, fat, lymph and blood vessels, and nerves [2].

In Sweden, around 250 people develop soft tissue sarcoma every year [3]. In the US, it is about 13 400 people every year [4]. These numbers indicate that soft tissue sarcoma is relatively rare worldwide. However, the survival rate for soft tissue sarcoma is lower than for many other cancer types. The 5-year survival rate explains what percent of people live at least 5 years after the cancer is found. For sarcoma, it is 65%, which is relatively low compared with, for instance, melanoma that has a 5-year survival rate of 94%. [5]

Handed in June 4, 2023
E-mail:{alma.lennartsson@gmail.com, emmafriberg10@gmail.com}
Technical supervisor: Caroline Brorsson, Qlucore
Swedish title: ”Hur molekylär diagnostik baserat på genuttryck kan förbättra diagnostiseringen av Sarkom”

B. Biology of sarcoma

The molecular biology of sarcomas is complicated. This is mostly due to the very heterogeneous nature of sarcoma tumors. The heterogeneity implies that all tumours differ, meaning that the phenotype and the genetic type is unique in every tumour type. Differences occur already during tumour development where complex biological processes affect the outcome [1].

These biological differences affect the ability to divide the tumours into subtypes. A subtype is a smaller group within a type of cancer, where the tumours have similar characteristics [6]. Sarcoma subtypes are determined based on how the tumour cells look under a microscope and are often classified based on where in the body the cancer began. However, soft tissue sarcoma tumors can involve multiple types of body tissues and sometimes their exact origin is unclear. [7].

There are over 80 subtypes of soft tissue sarcoma [7]. The most common ones in adults are liposarcoma (LPS), leiomyosarcoma (LMS) and undifferentiated pleomorphic sarcoma (UPS) [4]. LPS and LMS are located in fat tissue respectively smooth muscles, while UPS can be found anywhere in the body [7].

Since the biology of sarcoma tumours are very heterogeneous, the tumours within a certain subtype can differ as well. For instance, the subtypes liposarcoma (LPS) and rhabdomyosarcoma (RMS) can each be divided into four respectively five subgroups of their own [8].

C. Diagnostics

The low survival rate for patients with soft tissue sarcoma can partly be explained by the poor diagnostics. It is hard to determine which subtype of sarcoma a patient has, both due to the biology of the sarcoma tumour being complicated, as well as the lack of experience from physicians. Not being able to correctly diagnose the tumour leads to inadequate treatment for the patient [1]. Therefore, a solution to improve the sarcoma diagnostics is needed.

Few cases and opportunities to work with sarcoma leads to a lack of extensive knowledge and expertise among medical professionals, making accurate diagnosis more difficult. This emphasizes the importance of multidisciplinary collaboration

Table I
GLOSSARY

Subtype	A group within a type of cancer that share similar characteristics. [6]
RNA-sequencing	A way to measure the amount and types of RNA molecules in a sample, which can help understand how genes are expressed and regulated. [9]
Classifier	Supervised machine learning algorithm that learns to distinguish between different groups of data by identifying patterns, and can predict which group new data belongs to based on those patterns. [10]
PCA (Principal Component Analysis)	A statistical method used to reduce the dimensions of a big data set so that it can be visualized and analyzed easier, without losing too much information. [11]
UMAP (Uniform Manifold Approximation and Projection)	Algorithm that helps reduce high-dimensional data to a simpler form while preserving important relationships between data points. It is often used to create visualizations of complex data that reveal patterns or clusters. [11]
kNN (k-Nearest-Neighbours)	Classification algorithm that assigns a datapoint its label based on the label of the nearest neighbours/datapoints. The value of k, representing the number of neighbours used in the classification, is a hyperparameter that can be changed to optimize the accuracy. [10]
SVM (Support Vector Machine)	Classification algorithm that finds the optimal hyperplane to separate different groups of data in a high-dimensional feature space. The goal is to maximize the margin between the groups and separate them as accurately as possible. [10]
Random Forest	Classification algorithm that creates multiple decision trees on different random subsets of data and features, and combines their predictions to make more accurate predictions. It's useful when dealing with complex data or when overfitting is a concern. [10]
Cross-validation	A way to test how well a machine learning model can predict new data by splitting the available data into parts, training the model on some parts, and testing it on other parts. This helps to ensure that the model is not just memorizing the training data, but can generalize to new data. [12]
Confusion matrix	Table that summarizes the performance of a classification model by showing the number of correct and incorrect predictions for each class. [13]
Biomarker	Measurable characteristic or molecule that is used as an indicator of normal biological processes, pathogenic processes, or response to treatment. [14]

and expertise in sarcoma diagnosis. In Sweden, 5 different specialized sarcoma centers are responsible for all sarcoma diagnostics and treatments. These centers are located in Gothenburg, Linköping, Lund, Stockholm and Umeå [3].

A basic investigation of soft tissue sarcoma include the following steps:

- physical examination,
- tissue-based diagnostics (biopsy) that should be carried out at a sarcoma center,
- MRI (Magnetic Resonance Imaging) of tumour site,
- CT (Computer Tomography) of thorax [3].

These diagnostic methods are extensive, seen to both money, time and resources.

D. Treatment

Treatment options and chance of recovery are dependent on several different factors;

- the subtype of soft tissue sarcoma,
- size, grade and stage of tumour,
- where the tumour is located in the body,
- whether the whole tumour can be removed with surgery,
- age and general health,
- whether the cancer is recurring [15].

The stage of the tumour describes whether the cancer has spread to other parts of the body besides the soft tissue in which it originated. The staging process includes a series of tests and procedures. How quickly the tumour is likely to grow and how abnormal the cancer cells look under a microscope determines the grade of the tumour. All the gathered information determines the stage of the disease, which in turn is important for treatment planning. [15]

Surgery is the prime treatment option since few soft tissue sarcomas are curable with solely radiation- or chemotherapy. [3]. However, the surgical resection is often accompanied by chemotherapy and/or irradiation. The chemotherapy drugs could be administrated either before the surgery, to help reduce the size of the tumour, or after, to destroy remaining cancer cells. It could also be a combination of both. The optimal choice of treatment is largely dependent on the sarcoma subtype [1]. This highlights the need to diagnose patients with the correct subtype. It also allows the doctors to personalize the treatment. [7]

Personalized medicine, or precision medicine, is an advancing practice of medicine that uses the patients genetic profile to guide decisions regarding prevention, diagnosis and treatment of disease. By increasing the accuracy of diagnosis and identifying the most effective treatment therapy, the prognosis will increase as well [16]. Focus in this report is on how to improve the diagnostics of sarcoma, with a particular focus on the significant role of molecular diagnostics in achieving this goal.

E. Molecular diagnostics

Over the last decade, molecular diagnostics have developed significantly. Molecular diagnostics involves using gene expressions or DNA analysis to diagnose diseases. In this study, the main focus is on the usage of gene expressions. There is an important difference between gene expressions and whole genome DNA analysis. The gene expressions corresponds to the active processes in the body, since it represents the coding of the proteins in real time. Whereas a whole genome analysis only gives information of the DNA, but not about what genes are activated. Accordingly, molecular diagnostics of gene expressions is proved to be very efficient since a specific gene expression normally can be connected to a specific disease, which helps diagnosing correctly [17]. When analysing gene expressions in cancer, the tumour cells are examined. There are various ways to do this. Two commonly

used approaches are by using either DNA-methylation or RNA-sequencing (Table I), which are two separate methods to study gene expressions. [18]

Using RNA-sequencing within molecular diagnostics actually refers to observing the mRNA in the tumour. The mRNA, also known as messenger RNA, is important in the process of producing proteins. It is transcribed from the DNA in the cell nucleus and then transported to the cytoplasm where the protein synthesis takes place. Since an mRNA-strand corresponds to a specific protein, an analysis of the mRNA-strand will provide information about the real time processes in the cell [9]. In summary, by taking a sample of the tumours mRNA, it is possible to determine what processes are taking place in the tumour, which can be linked to certain diseases.

Within soft tissue sarcoma, studies have shown that the different subtypes differ in gene expressions, making it possible to distinguish a subtype based on the observed mRNA-strands [18]. This theory extends to other cancer types as well and has been used in several research studies to improve the cancer diagnostics [19].

The gene expression data can either be collected from clinical studies or from databases. The data will either way include a large amount of samples, where each sample corresponds to a patient. When using RNA-sequence data, every sample includes all the gene expressions found in the tumour of the patient. Moreover, every sample is also labeled with information about the patient, such as their sarcoma subtype, age, gender, etc. All this information can then be used to analyze the data. A large dataset is preferable since more samples correspond to more information. Therefore, many databases combine several different datasets to create a larger dataset. The datasets could, for example, come from different clinical studies, hospitals or patient groups. In this case, it is important to take into account that there might be differences between the data that could affect the outcome.

F. Data analysis

The second step in molecular diagnostics is the data analysis, which can include both visualization with various tools and diagrams, as well as more concrete machine learning algorithms, such as classification (Table I). For soft tissue sarcoma, it is of interest to determine which gene expressions correspond to which subtypes. To visualize this, methods such as PCA-plots and UMAP can be used (Table I). These methods will basically cluster the samples in different groups based on their gene expressions. Samples with more similar gene expressions will be located closer to each other in the plot. By coloring the samples based on their known subtype, potential patterns can be observed.

Furthermore, classifying the sarcoma data based on the subtype will provide more knowledge and can be used as a diagnostic tool. A classifier can be built in several different ways, depending on what algorithm is used. Common algorithms for building a classifier are Random Forest, k-Nearest Neighbours (kNN) and Support Vector Machine (SVM), see Table I [10].

In this case, the main function of the classifier would be to correctly assign an unknown tumour sample its sarcoma subtype. The accuracy of the classifier describes how well it pursues this task. For example, an accuracy of 0.9 means that 90% of the samples with unknown subtype were placed in their correct subgroup and 10% were classified incorrectly.

G. Similar studies

As mentioned, the goal with this report is to investigate how molecular diagnostics can help diagnosis of soft tissue sarcoma. Similar studies have already been made, both for soft tissue sarcoma and for other cancer types.

In a study from 2021, Koelsche et al. analyzed sarcoma data using clustering algorithms, which enabled identification of tumours sharing the same methylation pattern. Furthermore, a classifier was built with a Random Forest algorithm. The conclusion of the study was that molecular diagnostics can improve the diagnostics of sarcoma. In this case, by using DNA-methylation for the subtype classification [18].

Moreover, Qlucore have already used molecular diagnostics to classify different subtypes of leukemia [19]. This resulted in their newest software Qlucore Diagnostics [20], a tool to diagnose cancer based on their gene expressions. By using the patients data as an input, the program will determine which subtype of leukemia the patient has. The Qlucore Diagnostics software is still under development, with the goal to be used clinically in the near future. Meanwhile, Qlucore are developing similar classification tools for several other cancer types as well.

H. Purpose

The purpose of this study is to begin the process of building a diagnostic classifier for sarcoma. This will be done by constructing and then evaluating different classifiers. The main aspects that will be analyzed are the machine learning algorithms and the different subtypes. Since sarcoma diagnostics comes with many difficulties, it is crucial to do a detailed prestudy. Therefore, this study aims to collect information from previous studies, as well as generating new results to get one step closer to the implementation of a diagnostic classifier.

II. Method

The data used in this study is from “Treehouse Childhood Cancer Initiative at the UC Santa Cruz Genomics Institute” which consist of pediatric cancer genomic data [21]. Since this study cover soft tissue sarcoma, only the data labeled with sarcoma was collected for the definitive dataset. The dataset was downloaded as normalized log2 (TPM) and contained 826 samples and 58 581 number of variables.

The dataset was analyzed using a bioinformatic software called Qlucore Omics Explorer [22] which provides both visualization tools and ability to build and evaluate a classifier.

A. Data processing

The original dataset, without any alterations, was visualized in a PCA-plot, see Figure 1. This showed 3 samples that stood out compared to the others. A variance this large can for instance arise due to issues with the sample, issues in lab or technical problems when sequencing, but the details are unknown. Since this could affect the training of the classifier, these 3 samples were removed.

All sub-groups with 3 or less samples were then removed since they were assessed too few for training a classifier. The original dataset included the subtype osteosarcoma, which is a type of sarcoma found in bone, but since this study focuses on soft tissue sarcoma this group was removed as well. Ewing sarcoma can be found in both bone and soft tissue and was therefore maintained. The dataset also included heterogeneous subgroups of sarcoma that were ”not otherwise specified” and could not be used to build the classifier either. A new PCA-plot with the filtered dataset was then created, see Figure 2.

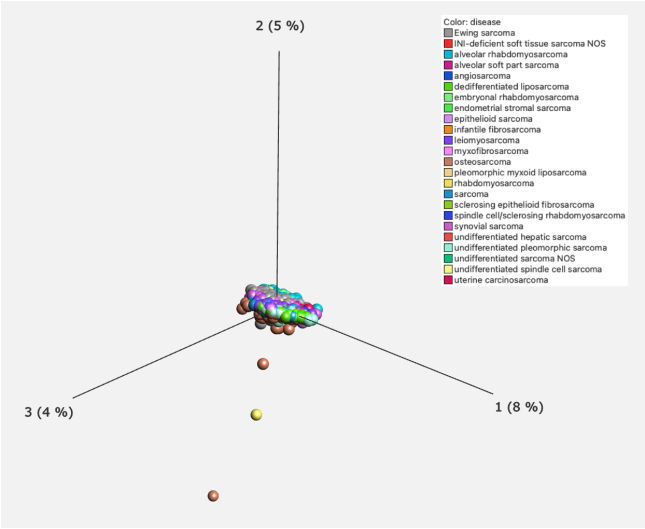

Figure 1. Principal Component Analysis (PCA) plot of gene expression data from soft tissue sarcoma samples. Each data point represents a sample, and their positions in the plot are determined by the underlying gene expression values. The axes of the plot correspond to the principal components, which are linear combinations of the gene expression variables that capture the most significant variation in the data. The colours of the samples correspond to their respective subtype. Original data is log2(TPM) with 826 samples and 58 581 number of variables.

B. Classification

The filtered dataset was then used to train classifiers with three different machine learning algorithms; kNN, SVM and Random forest (Table I). This was done in the Qlucore Omics Explorer software. Since the algorithms differ, the result might be affected depending on which one is used. Therefore, the algorithms were compared and the one with the best accuracy was chosen for further analysis. In this case it was the kNN algorithm with a k-value of 4 that presented the highest overall accuracy.

A classifier was then built in Python with a kNN algorithm, using a k-value of 4 and cross-validation (Table I). The result

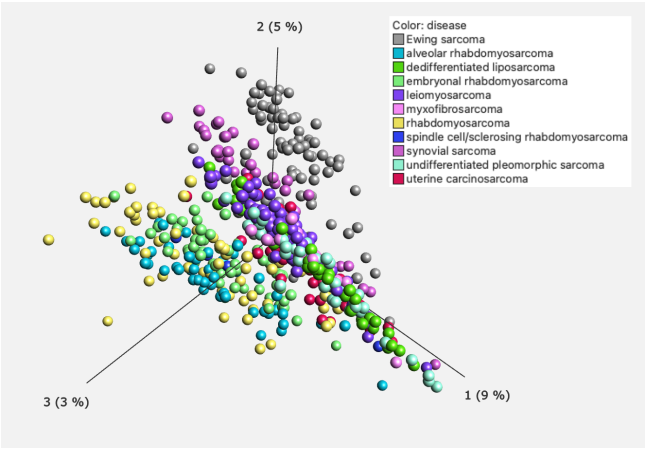

Figure 2. Principal Component Analysis (PCA) plot of gene expression data from soft tissue sarcoma samples. Each data point represents a sample, and their positions in the plot are determined by the underlying gene expression values. The axes of the plot correspond to the principal components, which are linear combinations of the gene expression variables that capture the most significant variation in the data. The colours of the samples correspond to their respective subtype. Filtered data is log2(TPM) with 546 samples and 58 581 number of variables.

was plotted in a confusion matrix (Table I), see Figure 3. The purpose of this was to enhance the accuracy of the classifier with the best performance. This was achieved by evaluating the classifier's performance on individual subtypes, which allowed for the exploration of different options for constructing the classifier and further improvement.

The confusion matrix showed that the classifier was not good at predicting the right label for the group rhabdomyosarcoma (RMS). It also showed that it often misplaced it with more specific types of RMS, mostly alveolar or embryonal RMS. Since the data set is a collection of data from different studies where different labels were used, it was concluded that the group labeled ”rhabdomyosarcoma” probably is a heterogeneous group containing more specific types of RMS.

The confusion matrix showed that the classifier was overall good at predicting the right label of both alveolar and embryonal RMS. However, not one sample of the spindle cell/sclerosing RMS was predicted correctly and was often confused with other RMS groups. This indicated that while all RMS subtypes may differ from the rest of the subtypes, it was still hard for the classifier to set them apart from each other.

Therefore, the heterogeneous RMS group was merged with the more specific types of RMS; alveolar-, embryonal- and spindle cell/sclerosing RMS. The aim with this was to enhance the accuracy of the main classifier comparing all of the subtypes to each other. A separate classifier containing only the three previously mentioned sub-groups of RMS was then trained and evaluated. The idea was to first use the main classifier to see if a sample belongs to RMS or not. If it does, the other classifier, specific to sub-groups of RMS, would give a new prediction of which type of RMS it is.

The final classifier, with the merged group of rhabdomyosarcoma, was built in Qlucore Omics Explorer with a kNN algorithm and a k-value of 4. Numerous k-values were evaluated

but the k-value of 4 resulted in the highest accuracy and was therefore chosen.

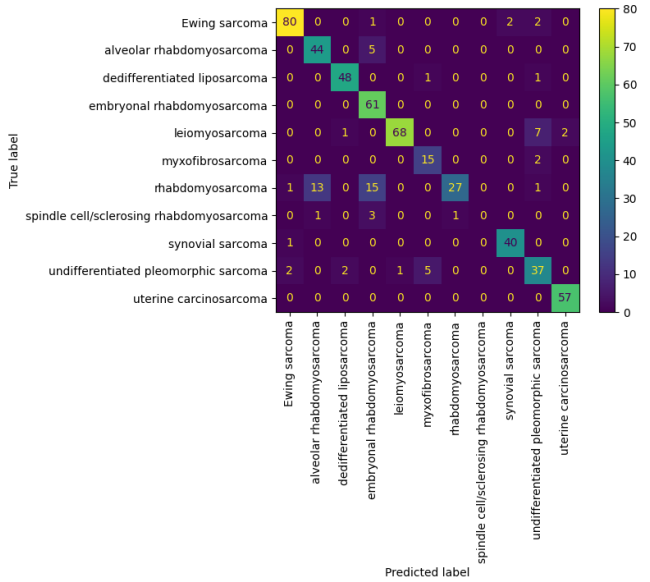

Figure 3. Confusion matrix that visualizes the performance of the classification algorithm. In this case the algorithm used was k-Nearest-Neighbour with a k-value of 4. Each row of the matrix represents the instances of the true class, while the columns represents the instances of the predicted class. The values following the diagonal shows how many samples that the classifier were able to place right, i.e. the predicted label is the same as the true label.

the groups of leiomyosarcoma (LMS), synovial sarcoma (SS), Ewing sarcoma (ES), rhabdomyosarcoma(all) (RMS) and uterine carcinosarcoma (UCS) are more distinctly clustered. Other groups, like myxofibrosarcoma (MFS), are harder to distinguish based on this plot. Overall, the UMAP gives an indication of which groups are easier to classify than others.

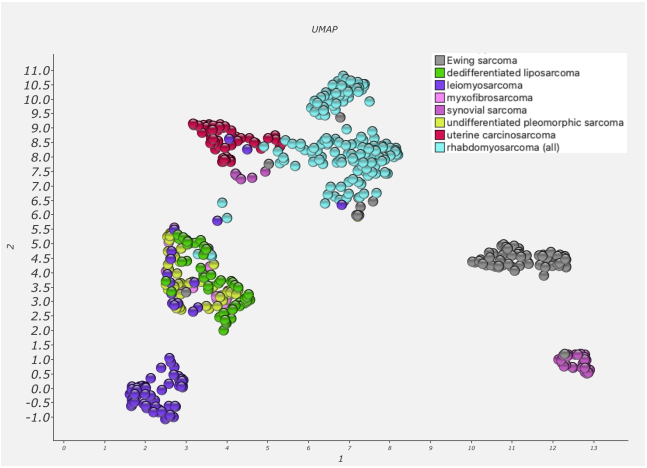

Figure 4. UMAP plot of soft tissue sarcoma gene expression data. The UMAP plot visually represents the high-dimensional gene expression data of soft tissue sarcoma samples in a two-dimensional space. Each data point in the plot represents a sample, and its position is determined by the similarity of its gene expression profile to other samples. The colour of the data points correspond to their respective subtype. The filtered data is log2(TPM) with 546 samples and 58 581 number of variables. Note that all of the RMS subgroups have been joined together to one, ”rhabdomyosarcoma (all)”.

C. Gene analysis

For further analysis, specific genes were more thoroughly examined. The goal of this analysis was to determine if a gene was significantly more expressed in a certain subtype than in others. If so, this gene could be a good biomarker (Table I) for the subtype.

Firstly, a two group comparison was made in Qlucore Omics Explorer. This means that the gene expressions for a specific subtype is compared against the gene expressions of all the other subtypes combined. Then, a statistical tool was used to determine which genes were significantly different between the two groups. By adjusting the significance level, it was possible to pin point the 20 genes that differed the most between the two groups. Lastly, some of these significant genes where analysed individually in a scatter plot. This procedure was repeated for several different subtypes and genes with the aim to distinguish which subtypes has good biomarkers and which does not.

One example of a subtype that was analyzed this way was leiomyosarcoma (LMS). By adjusting the significance, the gene MIR143HG was chosen for further analysis.

III. Results

A. Visualization

The processed data presented in a 2D-UMAP (Table I) is showed in Figure 4. This visualization shows that some subtypes are more easily distinguished than others. For instance,

B. Classification

The final kNN classifier resulted in an overall accuracy of 0.87, as seen in Table II. Table II also shows the specific accuracy for every subtype, which enables comparison between the subtypes. In this case, the majority of the subtypes have an accuracy above 0.8. Although, myxofibrosarcoma (MFS) and undifferentiated pleomorphic sarcoma (UPS) demonstrate a slightly lower accuracy of 0.64 and 0.60.

Table II
Evaluation of kNN classifier

Class	Accuracy	Number of samples
Ewing sarcoma (ES)	0.94	85
Dedifferentiated liposarcoma (DDLS)	0.98	50
Leiomyosarcoma (LMS)	0.81	78
Myxofibrosarcoma (MFS)	0.64	17
Rhabdomyosarcoma(all) (RMS)	1	56
Synovial sarcoma (SS)	1	41
Undifferentiated pleomorphic sarcoma (UPS)	0.60	47
Uterine carcinosarcoma (UCS)	0.98	57
Total	0.87	546

The separate kNN classifier, consisting only of the merged RMS groups, is presented in Table III. The overall accuracy was 0.60, however the groups vary in their individual accuracy.

Both alveolar and embryonal RMS has a relatively high accuracy compared to spindle cell/sclerosing RMS that has an accuracy of 0.

Table III
EVALUATION OF KNN CLASSIFIER FOR MERGED RMS SUBGROUPS

Class	Accuracy	Number of samples
Alveolar rhabdomyosarcoma (ARMS)	0.90	49
Embryonal rhabdomyosarcoma (ERMS)	0.94	61
Spindle cell/sclerosing rhabdomyosarcoma (SC-SRMS)	0	5
Total	0.60	115

Finally, the results from the other two machine learning algorithms is presented in Table IV and V. The kNN classifier has the highest total accuracy, as seen in Table II. Another observation is that the accuracy for MFS differs significantly between the algorithms. For Random forest, the accuracy is 0 and for SVM it is 0.18. Although, for the kNN classifier, the accuracy for MFS is much higher; 0.64. Additionally, the accuracy for UPS differs as well. It is higher for Random forest and SVM than it is for kNN.

Table IV
EVALUATION OF SVM CLASSIFIER

Class	Accuracy	Number of samples
Ewing sarcoma (ES)	0.95	85
Dedifferentiated liposarcoma (DDLS)	0.90	50
Leiomyosarcoma (LMS)	0.83	78
Myxofibrosarcoma (MFS)	0.18	17
Rhabdomyosarcoma(all) (RMS)	1	56
Synovial sarcoma (SS)	1	41
Undifferentiated pleomorphic sarcoma (UPS)	0.74	47
Uterine carcinosarcoma (UCS)	1	57
Total	0.83	546

Table V
EVALUATION OF RANDOM FOREST CLASSIFIER

Class	Accuracy	Number of samples
Ewing sarcoma (ES)	0.94	85
Dedifferentiated liposarcoma (DDLS)	0.90	50
Leiomyosarcoma (LMS)	0.85	78
Myxofibrosarcoma (MFS)	0	17
Rhabdomyosarcoma(all) (RMS)	1	56
Synovial sarcoma (SS)	1	41
Undifferentiated pleomorphic sarcoma (UPS)	0.89	47
Uterine carcinosarcoma (UCS)	1	57
Total	0.82	546

C. Gene analysis

An example of a gene expression connected to a certain subtype is visualized in Figure 5. In this case it is the gene MIR143HG which is significant for leiomyosarcoma (LMS). It is obvious that the gene expression levels are significantly higher for the samples within LMS than for the other subtypes. On the other hand, not all samples within LMS has a higher gene expressions since the field range over the lower values as well. Although, since the majority of the samples are higher, the overall conclusion is that MIR143HG might be a biomarker for LMS.

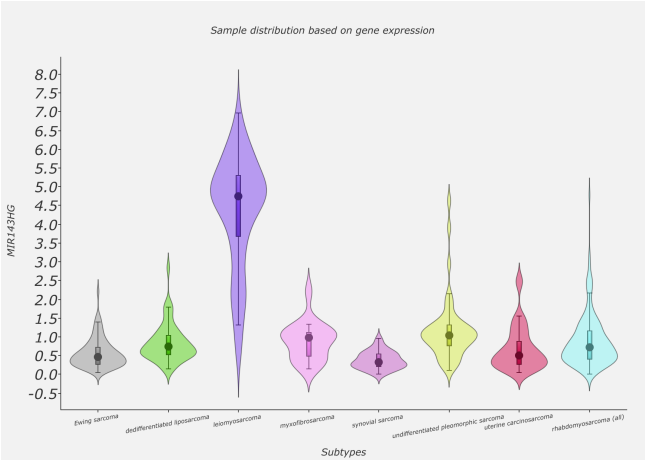

Figure 5. The violin plot displays how the gene expression levels vary for the samples within each subtype. The y-axis represents the amount of which the gene MIR143HG is expressed. The x-axis, and the different colored fields, represent one subtype each. The wider a colored field is at a specific place, the more samples contain that level of gene expression. Within the colored field, median, average value and standard deviation can also be observed.

IV. DISCUSSION

The final classifier resulted in an accuracy of 0.87 which is a relatively high number for a classifier. It means that if a patient was to be analysed, the chance of assigning the patient the correct subtype is 87%. However, the accuracy is not high enough to be used clinically, since health care requires more reliable diagnosing.

Another aspect is the width of subtypes in the classifier. As described, there are over 80 subtypes of soft tissue sarcoma, although this study only handled 10 of them. This means that only patients with one of these 10 subtypes could actually get diagnosed using the classifier.

On the other hand, even though the classifier was only able to classify 10 out of all the discovered subtypes of sarcoma, it is overall good at separating the most common soft tissue sarcomas from each other. This is under the assumption that the dataset represents the real clinical prevalence of different subtypes. For instance, the classifier contains some of the most common subtypes; liposarcoma (LPS), undifferentiated pleomorphic sarcoma (UPS) and leiomyosarcoma (LMS). Therefore, the classifier could potentially be a first assessment of the patient and if the tumour can not be classified, more tests will have to be done.

Even though the conclusion is that the classifier presented in this study is not good enough to be used clinically, there are many ways to improve it. A bigger dataset could improve both the width of subtypes and the accuracy of the classifier.

As seen in Table II and III, the subtypes vary in their individual accuracy. The subtype with the lowest accuracy, namely 0, was spindle cell/sclerosing RMS. There were only 5 samples with this label, and it is possible that the accuracy could be higher with an increased number of samples. The more samples, the easier to find patterns amongst them.

However, there are also subtypes with larger amount of samples that still has a low accuracy. For instance, both MFS and UPS have a relatively low accuracy. This proves that some subtypes might be more complicated to identify, based on their gene expressions. By observing the confusion matrix in Figure 3, it can be seen that many samples from MFS gets assigned UPS instead and vice versa. MFS and UPS are known to be very alike due to their similar molecular landscape [23]. Therefore it can be difficult to distinguish them from each other [24] [17].

Additionally, both MFS and UPS are represented by complex genomics, which also could explain the complications with the classification of these subtypes [25]. The Cancer Genome Atlas Research Network concluded in a study that the amount of myxoid stroma is what mainly separates UPS and MFS from each other. Therefore, analysing which genes are connected to the myxoid stroma levels and then use these specific genes to separate/classify MFS and UPS, was suggested as an option [17].

Another interesting result is how the different machine learning algorithms differ in accuracy, especially for some specific subtypes, see Table II IV V. As mentioned, the biggest difference is for the groups of MFS and UPS. A possible explanation for this is overfitting. Overfitting in a classifier occurs when the model becomes too complex and fits the training data too closely, leading to poor performance on new, unseen data. Essentially, the classifier is memorizing the training data instead of learning general patterns from it. [12] This will result in a high accuracy for the training data, but will not be representative for other data. A theory is that the kNN classifier is overfitted for MFS, since the accuracy is so much higher for kNN than for SVM or Random forest. To avoid this, bigger datasets could be used for the training of the classifier and it would be beneficial to use an independent validation dataset.

To summarize, the choice of algorithm impacts the result. Not only due to the chance of overfitting, but also in terms of the classifier outcome in general. It is of interest to use an algorithm that generates a high and reliable accuracy. Therefore it is crucial to test and evaluate all algorithms before choosing which one to use. Furthermore, adjusting the hyperparameters within each algorithm is another way of optimizing the classifier. For example, the k-value in the kNN algorithm can be varied to reach the best possible result.

A. *Further development of classifier*

A version of the classifier could, as discussed, be to first determine if the sample belongs to rhabdomyosarcoma (RMS) or not. If it does, the next layer of the classifier would predict which RMS subgroup the sample belongs to. This method could apply to myxofibrosarcoma (MFS) and undifferentiated pleomorphic sarcoma (UPS) as well; first determine if a sample belongs to MFS/UPS and then distinguish MFS and UPS from each other, for instance by analysing the genes connected to the myxoid stroma levels.

This could be extended even more. When building a classifier in Qlucore Omics Explorer, the program automatically uses Multi-Group comparison, also called ”all vs all”. This means that each group is compared to every other group in the dataset. This approach is useful when studying complex relationships or when there are multiple groups of interest that need to be compared simultaneously. [22]

To reach a higher classification accuracy a two-group comparison, ”one vs all”, could be used instead. This basically means that the classifier compares one group against all other groups combined. It simplifies the analysis by treating it as a binary problem: determining if the sample belongs to a particular group or not. This could then be repeated for every subtype, thus creating a sort of hierarchical classifier tree. Figure 6 displays a simple illustration of the hierarchical model in the form of a flowchart.

This type of layered classifier with two group comparisons is how Qlucore implemented their classifier for leukemia [20]. The challenge when implementing a similar model for sarcoma is mainly that there are extensively more subtypes to consider, which makes the hierarchical model more complex.

When implementing a hierarchical structure like this it is important for the system to be able to label a new sample as ”unclassified”. In this case, ”unclassified” means that the sample is not similar enough to the groups already included in the classifier. There are two main reasons for this. To start with, new subtypes are still being discovered, making it impossible to include them all in an established, pre-trained classifier. Additionally, even with a larger data set, it is not guaranteed that all subtypes can be accurately classified. To ensure that the final classifier is reliable and suitable for clinical use, it is better to prioritize higher accuracy in predicting the included subtypes rather than including more subtypes with lower accuracy.

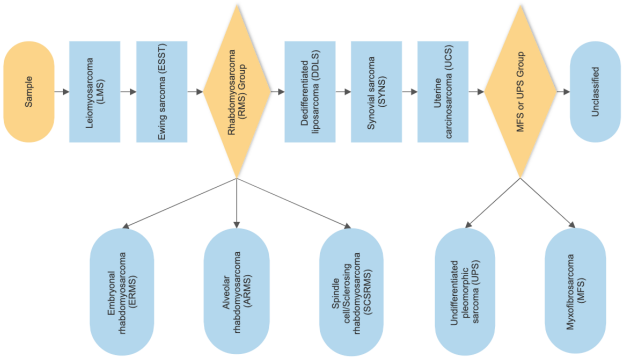

Figure 6. Flowchart of hierarchical classifier model. Note that the presented order of the subtypes does not hold any intentional significance.

B. Sustainable development

Sustainable development involves balancing economic growth, social progress, and environmental protection to create a more equitable and sustainable world. Health is essential for sustainable development as it is a basic human right and a significant contributor to economic growth. It does not only improve productivity and reduce healthcare costs, it also promotes social integration. [26]

A further developed classifier for soft tissue sarcoma subtypes as aid in diagnosis can contribute to sustainable development in several ways.

First and foremost, it can lead to more accurate diagnoses, leading to better treatment plans and reducing the risk of misdiagnosis or delayed treatment. This can improve patient outcomes, reduce healthcare costs associated with ineffective treatments, and minimize the burden on healthcare systems.

Furthermore, an improved classifier can aid in the development of personalized medicine approaches, tailoring treatment strategies based on the specific subtype of soft tissue sarcoma. This targeted therapy can potentially lead to better treatment responses, fewer side effects, and more efficient use of healthcare resources [16].

C. Ethics

The ethical concerns regarding a further developed classifier for soft tissue sarcoma include privacy protection, potential biases, accessibility and affordability, and the need for clear communication and informed consent.

Accurate classification may require access to sensitive personal health information. As always when handling patient data, there is a risk of privacy and confidentiality breaches. Ensuring data protection measures and receiving informed consent from patients becomes crucial to protect their privacy rights.

There is also a potential for bias and discrimination if the classifier is not validated over diverse populations. Biases in data collection, representation, or algorithmic decision-making can lead to differences in healthcare outcomes. Additionally, the cost and accessibility of using the classifier need to be considered. If the tool is expensive or inaccessible to certain populations or regions, it could worsen existing health inequities.

Clear communication and transparency are crucial to make sure that patients and healthcare providers understand the limitations, uncertainties, and potential errors associated with the classifier.

V. Conclusions

The study findings suggest that molecular diagnostics combined with machine learning algorithms have promising potential for improving sarcoma diagnosis. However, some subtypes are more difficult to distinguish and require more complex classification methods, such as hierarchical classifiers. The results also highlight the importance of expanding the dataset, including a wider range of sarcoma subtypes and optimizing the algorithm for better accuracy.

VI. Acknowledgements

We would like to express our deep gratitude towards our supervisor, Caroline Brorsson, at Qlucore, for amazing support and assistance throughout the project. We also want to thank Erik Hartman for his valuable guidance and expertise in both programming and machine learning, as well as Jan Nilsson for his assistance with Qlucore Omics Explorer. Finally, we would like to thank Carl-Johan Ivarsson who gave us the opportunity to do this project.

The workload was shared equally between the authors and both the data analysis and writing was performed together.

Appendix

References

[1] Grünewald TG, Alonso M, Avnet S, et.al"Sarcoma treatment in the era of molecular diagnostics," *EMBO Mol Med.* 2020 Nov 6;12(11):e11131. doi: 10.15252/emmm.201911131. Epub 2020 Oct 13. PMID: 33047515; PMCID: PMC7645378.

[2] National Cancer Institute, "Soft Tissue Sarcoma—Patient Version," 2023. [Online]. Available: https://www.cancer.gov/types/soft-tissue-sarcoma.

[3] Kunskapsbanken, "Nationellt vårdprogram skelett- och mjukdelssarkom i extremiteter och bålvägg," [Online]. Available: https://kunskapsbanken.cancercentrum.se/diagnoser/sarkom/vardprogram/.

[4] American Cancer Society, "Key statistics for soft tissue sarcoma," 2023. [Online]. Available: https://www.cancer.org/cancer/soft-tissue-sarcoma/about/key-statistics.html.

[5] American Society of Clinical Oncology, "Sarcomas, Soft Tissue: Statistics," 2023. [Online]. Available: https://www.cancer.net/cancer-types/sarcomas-soft-tissue/statistics.

[6] National Cancer Institute, "Dictionary of cancer terms; cancer subtype," 2023. [Online]. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cancer-subtype.

[7] Memorial Sloan Kettering Cancer Center "Types of Soft Tissue Sarcoma" 2023 https://www.mskcc.org/cancer-care/types/soft-tissue-sarcoma/types

[8] R. Kundra, H. Zhang, R. Sheridan,"OncoTree: A Cancer Classification System for Precision Oncology" *JCO Clinical Cancer Informatics* http://oncotree.mskcc.org

[9] National Human Genome Research Institute, "Messenger RNA (MRNA)," 2023. [Online]. Available: https://www.genome.gov/genetics-glossary/messenger-rna

[10] Towards Data Science "Top 6 Machine Learning Algorithms for Classification" 2023 https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

[11] Towards Data Science "Dimensionality Reduction for Data Visualization: PCA vs TSNE vs UMAP vs LDA" 2023 https://towardsdatascience.com/dimensionality-reduction-for-data-visualization-pca-vs-tsne-vs-umap-be4aa7b1cb29

[12] Amazon Web Services "What is Overfitting?" 2023 https://aws.amazon.com/what-is/overfitting/

[13] J. Mohajon "Confusion Matrix for Your Multi-Class Machine Learning Model" 2020 https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

[14] K. Strimbu, JA. Tavel "What are biomarkers?" *Curr Opin HIV AIDS.* 2010 Nov;5(6):463-6. doi: 10.1097/COH.0b013e32833ed177. PMID: 20978388; PMCID: PMC3078627.

[15] Dana-Farber Cancer Institute "Soft Tissue Sarcoma Overview" 2023 https://www.dana-farber.org/soft-tissue-sarcoma/

[16] National Human Genome Research Institute "Personalized Medicine" 2023 https://www.genome.gov/genetics-glossary/Personalized-Medicine

[17] The Cancer Genome Atlas Research Network "Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas" *Cell* vol 171, pp 950-965, 2017

[18] C. Koelsche, D. Schrimpf, D. Stichel, et al. "Sarcoma classification by DNA methylation profiling," *Nat Commun*, vol. 12, no. 498, 2021

[19] Qlucore "Acute Lymphoblastic Leukemia subtype classfication" 2023 https://qlucore.com/all-leukemia-diagnosis

[20] Qlucore "Qlucore Diagnostica" 2023 https://qlucore.com/diagnostics

[21] The Treehouse Childhood Cancer Initiative, “UCSC Treehouse Public Data,” 2023 [Online]. Available: https://treehousegenomics.soe.ucsc.edu/public-data/.
[22] Qlucore ”Qlucore Omics Explorer” 2023 https://qlucore.com/omics-explorer
[23] S. Haitao, L. Jilu, H. Fangyuan, et.al “Current research and management of undifferentiated pleomorphic sarcoma/myofibrosarcoma” *Frontiers in Genetics*, vol. 14, 2023.
[24] M. Yoshimoto, Y. Yamada, S. Ishihara,“Comparative Study of Myxofibrosarcoma With Undifferentiated Pleomorphic Sarcoma: Histopathologic and Clinicopathologic Review” *Am J Surg Pathol*,2020 Jan;44(1):87-97. doi: 10.1097/PAS.0000000000001389. PMID: 31651522.
[25] B. C. Widemann and A. Italiano ,“Biology and Management of Undifferentiated Pleomorphic Sarcoma, Myxofibrosarcoma, and Malignant Peripheral Nerve Sheath Tumors: State of the Art and Perspectives” *J Clin Oncol* vol 36, no 2, pp 160-167, 2018
[26] Sustainable Development Solutions Network ”Health in the Framework of Sustainable Development” 2014 https://resources.unsdsn.org/health-in-the-framework-of-sustainable-development