# Intrinsically Interpretable Machine Learning In Computer Aided Diagnosis

S. Ghosh, Sreejita

*DOI:*
10.33612/diss.175627883

# Chapter 8

# Conclusion

## 8.1 Outlook

This thesis introduces a set of intrinsically interpretable classifiers which not just achieves high performance metrics, but can also explain why a certain decision was made by the classifier. Interpretability of algorithms is extremely important when applied on anthropocentric applications, such as healthcare. An opaque machine learning model might achieve high performance metric even by learning *incorrect* features and might cause biased decision. (Obermeyer et al. 2019) reports a horrific ramification of blindly trusting an AI algorithm applied on healthcare setting in the United States. The decisions made by this particular algorithm were racially biased against treatment opportunities for the African-American community. In order to prevent such outcomes we need to focus on interpretability in addition to performance metrics of an algorithm.

The intrinsically interpretable classifiers introduced in this thesis are the angle-based variants of some of the classifiers of the LVQ family. These new variants can handle missing values of all types much more robustly than their Euclidean distance based predecessors, thus enabling learning from variable dimensional space. Additionally a geodesic averaging technique of LVQ models is introduced which makes it possible to preserve the model interpretability aspect of LVQ while combining it with the power of multiple learners, thus improving the model's performance. Finally we also introduced two probabilistic versions of the newly introduced angle-based LVQ and a probabilistic variant of the popular GMLVQ. All three of these classifiers use KL divergence as cost function during learning.

Chapter 2 introduces the concepts of interpretability, model transparency, and explainability, and explained the need for XAI in anthropocentric applications. Next it describes the types of interpretability and levels of transparency. Following that some of standard shallow ML classifiers are discussed in the context of explainability, interpretability, and transparency. Finally it discusses how model performance is often compromised for interpretability, and vice versa, possible model-agnostic solutions for minimizing this trade-off, and the limitations of these solutions.

Chapter 3 explains the main challenges of biomedical datasets which often hinder the direct application of the available standard ML techniques. These challenges included missing data, imbalanced classes, and heterogeneous measurements. Next it elaborates some of the available solutions for handling these issues and discussed their limitations, and thus provided the technical motivation for the angle based LVQ variants developed.

Chapter 4 mainly provides the biological motivation behind the angle based variants of LVQ introduced in this thesis. A real-life medical dataset containing all of the issues described in Chapter 3 is described. This dataset contained a publicly available synthetic dataset is provided on which interested readers can repeat the experiments in the thesis.

Chapter 5 introduces the LVQ variant capable of learning from variable dimensional space, such as data with missing values. This chapter also studied the effect of amount and type of missingness and amount of available training data on the dissimilarity measure used. Additionally it compared GMLVQ and Angle GMLVQ to state of the art shallow ML techniques such as KNN, RF, and LDA. It also introduced a geodesic averaging technique which is capable of combining the inherent interpretability of LVQ classifiers with power of ensembling. Next it illustrated the type of knowledge which can be extracted from real-life dataset.

In Chapter 6 the angle based variants of Local GMLVQ and Local LiRaM LVQ (also known as LVQ with two-matrix decomposition) are described. These variants are particularly useful for data with missing values and with decision boundaries which are non-linear. Due to local metric tensors in addition to being able to learn from more complicated datasets (wrt decision boundaries) these are able to extract condition specific information which could help stakeholders and end-users have better understanding about these conditions. Knowledge extracted from a publicly available dataset showed why it is a particularly difficult dataset when investigated as a five class problem. Clinical information extracted from the GCMS dataset helped our medical collaborators understand anomalies in the steroidogenic pathway better. The models being explainable were able to gain the trust of the clinicians as they could verify the biological findings by our presented classifiers.

Finally in Chapter 7 four variants of probabilistic LVQ classifiers are introduced, which use inclusive and exclusive KL divergence as cost function. Current investigation showed that of these four only the probabilistic variants of (i) GMLVQ using inclusive KL divergence ($P_I LVQ^E$), (ii) ALVQ using inclusive KL divergence ($P_I LVQ^A$), and (iii) ALVQ using exclusive KL divergence ($P_E LVQ^A$) are stable throughout the learning process. Our studies showed that in the presence of missing values $P_E LVQ^A$ is the most stable variant. These classifiers were also applied on the synthetic dataset and the GCMS dataset described in Chapter 4.

This thesis is titled *Intrisically Interpretable Machine Learning In Computer Aided Diagnosis* because the introduced set of models are interpretable by themselves. This means that no additional model-agnostic methods are needed to be applied on these models to extract explanation about why a certain decision was arrived at, or to visualize the decision boundaries, or to extract a representative example of each class that the model was trained on. These models performed feature selection in the input space and in an embedded way by learning the relevance of each of the features during training. The introduced models satisfied all the three criteria for intrinsic interpretability as defined by Backhaus and Seiffert in (Backhaus and Seiffert 2014), (Bibal and Frénay 2016), thus justifying the title of the thesis.

## 8.2 Future Work

Among future research which could be extended from this thesis we consider the following topics:

- The probabilistic ALVQ introduced in this thesis contains global metric tensor which is easily visualised, can learn from missing data and provide confidence of the classifier's decision. However the local metric tensors can extract a wealth of information about each condition the classifier is trained on. A probabilistic local ALVQ will combine all these properties.

- A systematic study of the introduced probabilistic LVQ classifiers in how well they are able to learn from noisy data, could enable comparison to probabilistic RF.

- Due to time complexity associated with the probabilistic GMLVQ using inclusive KL divergence the higher ranks of $\Omega$ when learning from the GCMS dataset could not be investigated. ALVQ with two matrices has similar issue. A short project aimed at optimising these classifiers for speed can greatly improve their application scope.

- The current version of probabilistic ALVQ and GMLVQ classifiers is capable of dealing with only one prototype per class. For more complex problems it would be useful to extend the current probabilistic LVQ variants introduced in this thesis to learn with multiple prototype per class.

- A comparative study of CE and both variants of KL divergence used as cost function for LVQ classifiers could help in better understanding of the nuances in these information theoretic principles.

## Closing remarks

(Ribeiro et al. 2016) reported that in a classification task to identify a wolf from a husky, even though the black-box model had high performance metrics, an XAI model on the black-box model revealed that the decision, whether wolf or husky, was made based on whether the background was snow. An alarmingly harmful effect of trusting in such black-box models is when applied in healthcare. (Obermeyer et al. 2019) mentioned how a black-box AI model predicted that African-American population, who were equally ill as the white people, were lesser at risk and therefore lesser likely for referral to programmes to take care of their complex medical needs. It further explained that these risk scores were assigned based on the total healthcare costs accrued in one year for the two demographic populations. While on one hand higher healthcare costs could suggest to higher health risks, alternatively, the low healthcare costs of the African-American population could be because on an average they could not afford high healthcare cost. The algorithm was clearly biased and therefore we need AL models or algorithms whose decisions are explainable, so that we can prevent such bias.

Even though the classifiers introduced in this thesis have been applied on medical datasets, these are applicable to similar datasets from other domains as well. Due to the recent ethical guidelines from the European Union (EU) for Trustworthy AI (Arrieta et al. 2020), more industries using ML are also very likely to look for *transparent* alternatives of ML. The work presented in this thesis will hopefully help the industries take a step closer towards making the transition to transparent and XAI. There are things in life which we have not yet found answers to, such as *How does pineapple go with pizza?*, or a relatively tolerable one, *Is Jaffa cake a cake or biscuit?*[42]. If some of us can seek answers to questions such as *How many of my scout cubs (8-10 years old) should I melt to get enough iron for a car?*[666], why should we stop asking an algorithm how it arrived to a particular decision and how sure it is of its decision? The non-technical aim of this thesis is to remind its few readers that if we start blindly trusting what we do not understand just because it is technical or looks fancy, or because it is becoming the *norm*, we will continue to see repetitions of the UK's GCSE results fiasco [2020].

[42] The more hilariously mind-blowing questions can be found in Randall Munroe's *What If*

[666] This really happened. It was asked by a school student, at an online science communication event I participated in.

[2020] More information in case you missed it can be found at https://en.wikipedia.org/wiki/2020_UK_GCSE_and_A-Level_grading_controversy