Early Mastitis Diagnosis through Topological Analysis of Biosignals from Low-Voltage Alternate Current Electrokinetics

Zhifei Zhang, Yang Song, Haochen Cui, Jayne Wu, Fernando Schwartz and Hairong Qi.

***Abstract*— Mastitis is the most economically important disease of dairy cows worldwide, and it constantly plagues the dairy industry. A reliable biosensing method is thus imperative to detect this disease at its early stage and accurately identify the pathogen concentration level in order to better control the disease and consequently improve the quality of milk. Recent research indicates that shorter assay time and/or higher sensitivity can be achieved by integrating alternate current electrokinetics (ACEK) with biosensing. However, most existing ACEK devices use voltage levels around 10V at the risk of electrochemical reactions because a lower voltage may not effectively trigger the ACEK effect. Currently, there are no related works that can efficiently tackle the dilemma between avoiding electrochemical reaction and accelerating assay process. This paper adopts low-voltage (40∼135mV) ACEK, which is safe but yields ambiguous biosignals within a short assay time, presenting great challenge to high-fidelity identification of pathogen concentration levels. This paper makes two distinctive contributions to the field of biosignal analysis. First, moving away from the traditional signal analysis in the time or spectral domain, we exploit the possibility of representing the biosignal through topological analysis that would reveal the intrinsic topological structure of point clouds generated from the biosignal. Second, in order to tackle another common challenge of biosignal analysis, i.e., limited sample size, we propose a so-called Gaussian-based decision tree (GDT), which can efficiently classify the biosignals even when the sample size is extremely small. Experimental results on the classification of five pathogen concentration levels using only 10 samples taken under various voltage levels demonstrate the robustness of the topological features as well as the advantage of GDT over some other conventional classifiers in handling small dataset. Our method reduces the voltage of ACEK to a safe level and still yields high-fidelity results in a short time.**

I. INTRODUCTION

Mastitis is caused by bacterial infection of udder tissues. Once the infection happens, the immune system of cows will respond and fight the infection with an increase in the number of immune cells, referred to as the somatic cells, primarily white blood cells. The number of somatic cells in milk, i.e., somatic cell count (SCC), is an important measure of milk quality used throughout the world. Milk with a high SCC is associated with a higher incidence of antibiotic residues in milk and the presence of pathogenic organisms and toxins in milk. Annually, mastitis causes approximately $2 billion in losses to the U.S. dairy industry, and 60% of the losses, or $1.2 billion, comes from subclinical mastitis.

Cows with clinical mastitis are easy to identify with the swollen teats and thick, curdled discharge in the milk, while subclinical cases, the most common form of mastitis, are difficult to detect due to no obvious clinical symptoms of the illness and no visible changes to the milk composition. Subclinical mastitis in cows can still lead to abnormally high SCC in milk and can be up to 40 times more common than clinical cases of the illness. The primary focus of most subclinical mastitis programs is to reduce the prevalence of the contagious pathogens, S agalactiae and S aureus, as well as other gram-positive cocci. Because mastitis can be caused by many different pathogens, the early availability of diagnostic results, i.e., the pathogens that cause an elevated SCC in a herd as well as their concentration levels, is crucial towards controlling the spread of new infections.

Alternating current electrokinetics (ACEK) [1], implemented by microelectrodes immersed in sample fluids, induces directional particle or fluid motion by externally applying AC electric field over the electrodes. Since its advent in the 1990s, ACEK has been widely studied and utilized to accelerate the movement of macromolecules towards sensing areas [2], [3]. Most existing ACEK devices use AC around 10V because lower voltage may not efficiently accelerate the assay process, which means we have to wait for tens of minutes or even hours before achieving an available result. However, applying a voltage higher than 1V over biofluids raises the risk of electrolysis, biofouling, etc. Currently, there are no related works that can efficiently tackle this dilemma.

This paper presents a topological signal processing scheme to achieve short assay time under low-voltage ACEK. This method allows much lower voltage (40∼135mV) for ACEK that significantly reduces the risk of electrochemical reaction. In addition, it still yields relatively high identification accuracy within a short assay time (30 seconds). To verify our method, milk samples with five concentrations of bovine IgG whole molecules (0μg/lm, 1μg/lm, 5μg/lm, 10μg/lm, 100μg/lm) are tested using exactly the same biosensor. We first extract robust and representative features using topological analysis of the obtained biosignals. Due to the limitation in the number of milk samples, only two biosignals are obtained from each concentration. Such a small sample size fails most conventional classifiers, such as SVM, kNN, decision tree, etc. Thus, the second contribution of the proposed work is a Gaussian-based decision tree that efficiently handles the small sample size problem.

The rest of this paper is organized as follows. Section II describes the challenges of extracting robust and representative features under low-voltage ACEK. The topological-based feature extraction and the Gaussian-based decision tree classifier are discussed in Sections III and IV, respectively.

Zhifei Zhang (zzhang61@vols.utk.edu), Yang Song, Haochen Cui, Jayne Wu, and Hairong Qi are with the Department of Electrical Engineering and Computer Science, Fernando Schwartz is with the Department of Mathematics, the University of Tennessee, Knoxville, TN 37996, USA.

Section V provides experimental study. Section VI concludes the paper.

II. CHALLENGES OF LOW-VOLTAGE ACEK

The biosignal obtained under low-voltage ACEK within short time duration (e.g., 30 sec) exhibits challenging characteristics that would largely hinder the early pathogen diagnosis process. The challenges are reflected from three perspectives, including small inter-class variation, large intra-class variation, and random oscillation.

As shown in Fig. 1(a), the biosignals of different concentrations are difficult to separate in the time domain due to the large overlap. In other words, the distance between samples of different classes (i.e., inter-class variation) can be quite small. On the other hand, it is also possible that two signals from the same class (i.e., intra-class variation) are significantly different from each other, as illustrated in Fig. 1(d). In addition, the random oscillation refers to the randomness inherent in both period and amplitude of the biosignal, as shown in Fig. 1(a, d), which will fail the class of spectral analysis methods, such as Fourier transform and discrete wavelet transform.

These unique challenges of biosignals obtained from low-voltage ACEK in short duration call for the design of robust and representative feature extraction approaches as well as effective classifiers that can cope with the small sample-size problem.

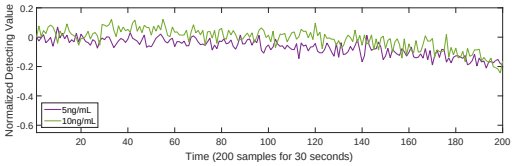
III. TOPOLOGICAL ANALYSIS FOR FEATURE EXTRACTION

As discussed above, robust and representative features are difficult to obtain from the time and/or spectral domain. On the other hand, topological analysis transforms the biosignal into higher-dimensional topological space, where the biosignal is represented as point cloud, using which more separable features can be extracted by analyzing the shape and intrinsic structure of the cloud. For example, biosignals with small inter-class variation (Fig. 1(a)) show larger variation in the shape of point cloud (Fig. 1(b, c)), especially the length of the major axis of the fitted ellipse. And biosignals with large intra-class variation (Fig. 1(d)) share similar length of the major axis of the fitted ellipse, as shown in Fig. 1(e, f).

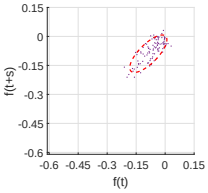
The delay embedding (DE), a powerful tool of geometric time series analysis [4], is applied here to convert the biosignal to 2D topological space. Suppose a signal sequence can be represented by a discrete function $f(x),\ x \in \mathbb{Z}^+$. Given a delay step size $s \in \mathbb{Z}^+$ and a target dimension $d \in \mathbb{Z}^+$, the DE of $f(x)$ at $t \in \mathbb{Z}^+$ is expressed in Eq. 1.

$$DE(f,t;s,d) = \begin{bmatrix} f(t) \\ f(t+s) \\ \vdots \\ f\left(t+(d-1)s\right) \end{bmatrix} \tag{1}$$

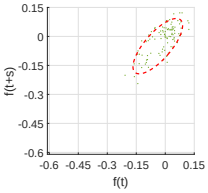
Assume there are n sampling points in the signal sequence. If sliding t from the first to the last available point, we will obtain $m = n-(d-1)s$ d-dimensional points, which form the point cloud in the topological space. Fig. 2 illustrates



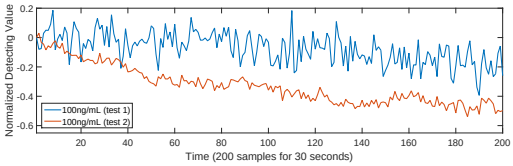
(a) Biosignals with small inter-class variation



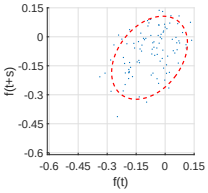
(b) Point cloud (5ng/mL)



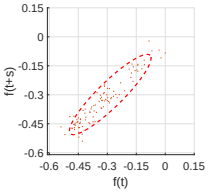
(c) Point cloud (10ng/mL)



(d) Biosignals with large intra-class variation



(e) Point cloud (test 1)



(f) Point cloud (test 2)

Fig. 1. Examples of biosignals and their representation in the topological space. (a) and (d) display biosignals under 40mV-ACEK. The denotations indicate corresponding concentrations of bovine IgG whole molecules in the test milk samples. In (b), (c), (e) and (f), point cloud (colored dots) are generated through delay embedding from the biosignals in corresponding colors. Red dash curves display fitted ellipses on the point cloud.

the procedure of DE. The left shows sampling points (red dots, $n = 5$) in the time domain. Let $s = 1$ and $d = 2$. When $t = 1$, DE yields a 2D point $(f(1), f(2))$ according to Eq. 1. The dots with green circles indicate the points used in DE ($t = 1$), and the corresponding 2D point is shown in the right with the same color. By the same token, when $t = 2$, the points with yellow circles are collected to generate the next 2D point $(f(2), f(3))$, which is shown as yellow dot in the right. Increasing t, we finally obtain a point cloud with $m = 4$ points.

From the point cloud, shape analysis and persistent homology [5] can then be performed in parallel, to extract shape features and intrinsic structure information, respectively. The shape analysis calculates a few parameters of the fitted ellipse, such as length of axes, area, orientation, etc. They represent the geometric characteristics of the point cloud. The intrinsic structure is investigated by persistent homology—a method for computing topological features of the point cloud at different spatial resolutions [6]. It is computed based on simplicial complex [7], which may

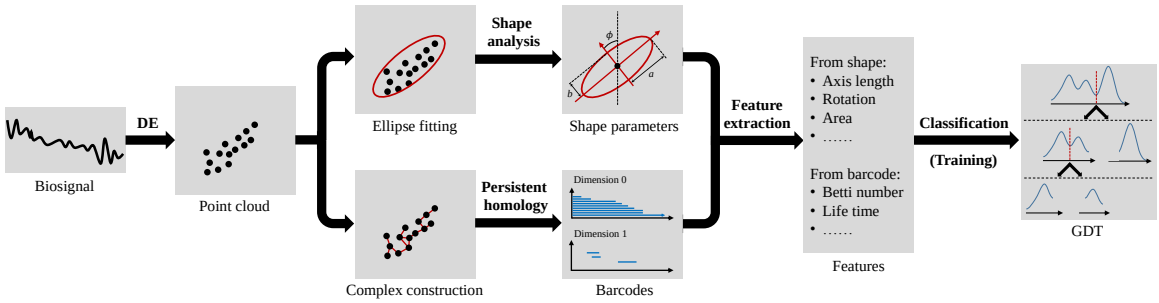


Fig. 3. The flow of topological analysis on the biosignal. The biosignal in time domain is converted to 2D point cloud in the topological space through DE. Then, the flow divides into two branches: upper branch (shape analysis) fits the point cloud with an optimal ellipse and extracts its parameters; lower branch (persistent homology) constructs complex and computes barcodes on it. Finally, features from upper and lower branches are merged to train the classifier GTD, which is designed to handle small dataset.

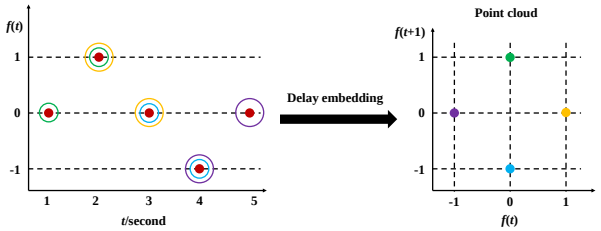


Fig. 2. Delay embedding ($s = 1$, $d = 2$). Left: red dots denote sampling points in the time domain, and circles in the same color indicate the point group used in DE. Right: point cloud in the 2D topological space, in which each dots is generated through DE from the sampling points marked by the same color circles.

be simply considered as a connected graph of the point cloud. Any two points, whose Euclidean distance is smaller than certain threshold ϵ, will be connected. Meanwhile, any convex hull constructed by points whose pair-wise distances are smaller than ϵ is filled to be a solid block. Fig. 4 illustrates the basic idea of persistent homology. With the increase of ϵ ($\epsilon_i < \epsilon_j, \forall i < j$), more points are connected, and more blocks are generated as shown by colored fields. The figure shows that five independent points are eventually connected to be a single component. Note that a "hole, i.e., the area is enveloped by connected point but not filled, formed at $\epsilon = \epsilon_3$, persists to $\epsilon = \epsilon_4$, but is filled (disappears) at $\epsilon = \epsilon_5$. Persistent homology tries to find persistent features during some dynamic processes.

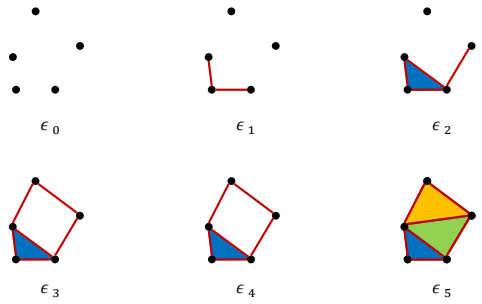


Fig. 4. Persistent homology on 2D point cloud, $\epsilon_i < \epsilon_j, \forall i < j$. With the increase of ϵ, more points are connected, and more areas are filled. At ϵ_3, a hole is formed, then the hole is filled (disappears) at ϵ_5. Thus, $\epsilon_5 - \epsilon_3$ indicates the lifetime of the hole.

Specifically, persistent homology tracks two dynamic processes: 1) how fast those points are connected to a single component and 2) when a hole appears and disappears. The holes are of paramount importance for discovering the intrinsic structure of a point cloud [5], [8], [9]. Therefore, in what we are interested is at what value of ϵ a hole appears and how long it persists until being filled. Assuming a hole appears at ϵ_{birth} and dies at ϵ_{death}, the length it persists is $\epsilon_{death} - \epsilon_{birth}$, which is referred to as the lifetime of the hole. Usually, lifetimes are visualized by barcodes, the ends of which denote the birth and death points of holes (dimension 1), as shown in Fig. 3. The barcodes (dimension 0) indicates the speed the points are connected to a single component. Generally speaking, the longer the barcode, the slower the connection speed.

The flow of feature extraction based on topological analysis is elaborated in Fig. 3.

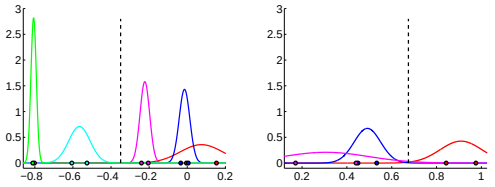
IV. GAUSSIAN-BASED DECISION TREE

Empirical experiments show that most conventional classifiers, such as SVM, decision tree, kNN, etc., fail to handle small datasets (e.g., only 2 samples for each class in our case). The Gaussian-based decision tree (GDT) is proposed to solve this problem. GDT assumes Gaussian distribution on each feature from samples of the same class, which is motivated by the fact that a feature from biosignals of the same class is supposed to vary in a finite interval. This assumption carries two advantages: 1) speeds up the learning procedure and 2) makes GDT less sensitive to the sample size.

During the construction of decision tree, given a feature $\mathbf{v}$ at certain node, the optimal splitting boundary can be obtained by Eq. 2. Assume there are $n \geq 2$ classes at the node. Means of each class are calculated on $\mathbf{v}$ and sorted in ascending order in $\boldsymbol{\mu} \in \mathbb{R}^n$. The corresponding standard deviation is stored in $\boldsymbol{\sigma} \in \mathbb{R}^n$.

$$\begin{aligned} \mathcal{F}(\mathbf{v}) = \min_{x_i} & \left\{ \frac{G_{\mu_i,\sigma_i}(x_i)}{\mu_{i+1} - \mu_i} + \gamma E(x_i) \right\} \\ \text{s.t. } & G_{\mu_i,\sigma_i}(x_i) = G_{\mu_{i+1},\sigma_{i+1}}(x_i) \\ & \mu_i < x_i < \mu_{i+1},\ i = 1, 2, \cdots, n-1 \\ & \mu_i \in \boldsymbol{\mu}, \sigma_i \in \boldsymbol{\sigma} \end{aligned} \tag{2}$$

where $G_{\mu_i,\sigma_i}(\cdot)$ denotes the Gaussian function whose mean and variance are μ_i and σ_i^2, respectively. γ adjusts the effect of entropy $E(x_i)$ given a boundary x_i. On the given feature $\mathbf{v}$, only the intersections of the closest Gaussian distributions are involved in searching the optimal boundary. In addition, features from each class are fitted to separated Gaussian distributions regardless of the sample size, which makes GDT more capable to handle small datasets. Iterating Eq. 2 on each feature at a node, the global optimal feature and splitting boundary will be obtained as shown in Fig. 5.



(a) Root node (b) Leaf node

Fig. 5. A sample of searching the optimal boundary. The dots in different colors denote samples from different classes, and the colored curves represent corresponding Gaussian distributions. The vertical dash line shows the optimal splitting boundaries. The horizon axis shows the value of certain feature. (a) is the root node, and (b) is a leaf node of (a). The horizon axis displays values of the selected feature, which are different in (a) and (b).

V. EXPERIMENTS AND RESULTS

In the experiment, the same biosensor is used to test the milk samples of five concentrations (classes) under two voltage levels (40mV and 135mV). The five concentrations correspond to different levels of pathogen, and the two voltages are used to illustrate the voltage effect on ACEK. For each concentration-voltage pair, only two biosignals are obtained through 30-second assay. Therefore, there are two samples for each five classes under certain voltage.

First, the biosignals are converted to 2D point cloud by DE. We do not convert the signal to higher dimension for two reasons: 1) higher dimension results in higher computation complexity, and 2) empirical experiment shows that higher dimension unnecessarily provides more redundant features. Following the flow shown in Fig. 3, features are listed in Table I.

TABLE I

FEATURES EXTRACTED FROM TOPOLOGICAL ANALYSIS

Denotation	Description
shape-a	Length of the major axis of the ellipse.
shape-b	Length of the minor axis of the ellipse.
shape-ϕ	Orientation of the ellipse.
shape-A	Area of the ellipse.
shape-R	Ratio of the major and minor axes of the ellipse.
homo-0	Accumulated lifetime of independent connected components. Integration of bars in the barcode (dimension 0).
homo-1	Accumulated lifetime of holes. Integration of bars in the barcode (dimension 1).

Statistical features in the time domain are also extracted for comparison purpose, including mean, standard error, skewness and kurtosis. Under each voltage, different classifiers are implemented on topological and statistical features, respectively. To avoid over-fitting, the leave-one-out cross validation is applied for each classifier. Table II displays the results. The accuracy of each classifier under the same voltage shows that topological features outperform statistical features, especially under lower voltage. As the voltage decreases, the accuracy of statistical features degrades dramatically. Instead, the topological features still preserve relatively high performance. Comparing the classifier, GDT achieves the best performance on the small dataset.

TABLE II

ACCURACY OF DIFFERENT CLASSIFIERS USING DIFFERENT FEATURES

Classifier	Statistical features		Topological features	
	40mV	135mV	40mV	135mV
GDT	60%	90%	80%	100%
decision tree	30%	70%	80%	70%
random forest	30%	70%	40%	70%
SVM	50%	50%	50%	60%

VI. CONCLUSIONS

This paper presented a topological-based method to achieve fast and high-fidelity biosensing under low-voltage ACEK. In the topological space, robust and representative features are extracted from the biosignals which appears indistinguishable in the time domain. To efficiently classify the quality level of milk using a small sample size, the Gaussian-based decision tree (GDT) is proposed, which outperforms most existing classifiers.

This method has the potential to analyze other biosignals that are not easy to be distinguished in time or spectral domain, due to its effectiveness in handling some challenging issues like significantly random oscillation, small inter-class variation, large intra-class variation, and extremely small sample size. The topological-based method has shown itself to be an effective solution to time serial biosignal analysis.

REFERENCES

[1] H. Morgan and N. G. Green, *AC electrokinetics: colloids and nanoparticles*. Research Studies Press, 2003, no. 2.

[2] J. Wu, "Biased ac electro-osmosis for on-chip bioparticle processing," *Nanotechnology, IEEE Transactions on*, vol. 5, no. 2, pp. 84–89, 2006.

[3] K. Yang and J. Wu, "Numerical study of in situ preconcentration for rapid and sensitive nanoparticle detection," *Biomicrofluidics*, vol. 4, no. 3, p. 034106, 2010.

[4] M. R. Muldoon, D. S. Broomhead, J. P. Huke, and R. Hegger, "Delay embedding in the presence of dynamical noise," *Dynamics and Stability of Systems*, vol. 13, no. 2, pp. 175–186, 1998.

[5] A. Zomorodian and G. Carlsson, "Computing persistent homology," *Discrete & Computational Geometry*, vol. 33, no. 2, pp. 249–274, 2005.

[6] G. Carlsson, "Topology and data," *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 255–308, 2009.

[7] L. J. Guibas and S. Y. Oudot, "Reconstruction using witness complexes," *Discrete & computational geometry*, vol. 40, no. 3, pp. 325–356, 2008.

[8] H. Edelsbrunner, "Persistent homology: theory and practice," 2014.

[9] X. Zhu, "Persistent homology: An introduction and a new text representation for natural language processing," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*. AAAI Press, 2013, pp. 1953–1959.