Signal Analysis Methods of Various Nanopore Sensors

A Project Submitted in Partial Fulfilment of the Requirements for the Degree of

BACHELOR OF TECHNOLOGY

IN

Electronics and Telecommunication Engineering

By

Rohit Ray

EXAM ROLL-510714044

Swapnil Roy

EXAM ROLL-510714008

Under The Guidance of

Dr.ChirasreeRoyChaudhuri

DEPARTMENT OF ELECTRONICS & TELECOMMUNICATION ENGINEERING

INDIAN INSTITUTE OF ENGINEERING SCIENCE & TECHNOLOGY, SHIBPUR

HOWRAH- 711103

MAY 2018

FORWARDING

I hereby forward the project entitled **"Signal Analysis Methods of Various Nanopore Sensors"** submitted by Rohit Ray (510714044) and Swapnil Roy (510714008) under my guidance and supervision in partial fulfilment of the requirements for the degree of **"Bachelor of Technology"** in **Electronics & Telecommunication Engineering** in the Department of Electronics & Telecommunication Engineering, Indian Institute of Engineering Science & Technology, Shibpur.

Dated:

……

(Dr.Chirasree RoyChaudhuri)

Assistant Professor

Department of Electronics & Telecommunication Engineering

COUNTERSIGNED BY:

……

(DR. MonojitMitra)

Professor& Head,

Department of Electronics & Telecommunication Engineering

CERTIFICATE OF APPROVAL

The foregoing project work is hereby approved as a creditable study of Engineering subject carried out and presented in a satisfactory manner to warrant its acceptance as a prerequisite for the Degree of **'Bachelor of Technology'** in the Department of Electronics & Telecommunication Engineering, Indian Institute of Engineering Science and Technology, Shibpur for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the report only for the purpose for which it is submitted.

Board of Thesis Examiners:

_____ _____

_____ _____

_____ _____

_____ _____

_____ _____

_____ _____

ACKNOWLEDGEMENT

We would like to thank <name> and <name>, research scholars, in the Department of Electronics and Telecommunication Engineering, for providing us with all the necessary data and amenities to complete our project.

We would also like to thank Department of Computer Science and Technology for providing us with high performance servers which were required for carrying out the heavy computations required.

Contents

Foreword

In this thesis we delve into various areas of estimation and quantification of Biological molecules through different types of sensors and data analysis methods.

This thesis explores different methods of data acquisition and analysis using artificial intelligence and various soft-computing models.

Chapter 1

Nanopore Sensors

In this thesis we delve into two different types of nanopore sensors, namely,

(i) Glass nanopore sensor
(ii) Nanoporous Graphene FET sensor

1.1 Glass Nanopore Sensor:

Solid-state nanopores for single-molecule detection of DNA were first used in 2001 by Li et al., who were able to detect double-stranded (ds) DNA when it was translocated through the nanopore. To fabricate this nanopore, an electron beam from a transmission electron microscope (TEM) was focused on a few nanometer thick silicon nitride membrane. The membrane separates two reservoirs containing a salt solution, across which an electric potential was applied. This causes an ionic current that is measured and depends on parameters such as salt concentration and the dimension of the nanopore. Nanopores are a useful technique to detect charged single molecules, which can be forced to translocate through the nanopore by the electric field. The volume displaced by the molecule can be sensed, because when it enters the nanopore, there is a reduction in the ionic current at high salt concentrations.

The ability to precisely control the size of the nanopore is crucial to sense small molecules since the amplitude of the conductance drop increases with smaller diameters. An early technique used the electron beam of an already drilled nanopore to shrink it to any desired diameter. The electron beam would heat up the silicon membrane around the nanopore, reducing the size of the pore due to the surface stress and the increased mobility of the heated atoms. The method of pore shrinking with a transmission electron microscope was soon expanded to other instruments such as scanning electron microscopes or lasers, which were able to locally heat up a material.

An inexpensive alternative to solid-state nanopores in silicon membranes is laser pulled **glass nanocapillaries**, which have a conical shape and a very small orifice at their tip. It was shown that they are able to sense the folding state of translocating dsDNA on the single-molecule level.

1.1.1 Working:

Nanopores have been established as a promising platform for detection of DNA, RNA, proteins and other molecules at single molecule level, stochastically in real time by ensemble averaging of the current blockade events. Further, the throughput of the acquired signal is independent of the sample volume and depends on the concentration which facilitates the integration of nanopore array with simple analyte delivery arrangement. Moreover, glass

nanopore array can be fabricated for simultaneous sensing of multiple molecules at a significantly low cost than the nanowire platforms [1]. Broadly, four approaches have been adopted for identification of proteins in complex analytes using these nanopores- firstly, discriminating the translocation time and current blockade of a specific receptor conjugated protein from that of the free protein or the receptor, secondly, the modification of a double stranded DNA with specific chemical motifs and identifying the molecule specific sub-threshold current peaks, thirdly, modifying the nanopore walls with specific ligands embedded mobile lipid bilayer membrane and detecting the target molecule from the translocation time and fourthly, functionalizing the pore walls with specific antibodies/aptamers and determining the target molecule from the amount of current blockade. The translocation speeds of most of the protein complexes are fast which makes it particularly challenging for quantification. The methodology using grafted ds-DNA is exceptionally useful but has been reported for quantification of target proteins with concentration in the range of few nanomolars. Attachment of lipid bilayer membrane is a novel approach towards detection of picomolar concentration of target proteins but quantification becomes erroneous if the non-target protein is present in quantities larger than 300 times. Moreover the technique requires critical surface engineering for tailoring the viscosity corresponding to the target molecule and hence repeatability is a major challenge. Further, the detection requires labeling of protein molecules. On the contrary, functionalization of nanopore walls with aptamers is a specific label free process, where the entry of each target protein blocks the current at different levels. This method has enabled the detection of proteins and small molecules in picomolar range in control solution but it's quantification in complex analyte down to few picomolar is still challenging during real time operation. Quantification is critical, especially near the lower threshold limit to diagnose whether the patient is progressing towards the diseased state.

1.1.2 Specifications of Sensors used:

Experiments have been performed with conical nanopores of diameters 30±3 nm for aptamer immobilization. The schematic representation, typical SEM and conductance measurements before and after functionalization are indicated in Fig.1.1(a-c). For sensing the presence of thrombin in buffer, it has been added at concentrations ranging from 50 pM to 500 pM in the glass capillary. The transpore current traces have been recorded with both positive and negative voltages inside the capillary with respect to the external solution. It has been observed from Fig.1d that hardly any thrombin molecule has reached the pore tip. In presence of positive voltage of + 250mV, few spikes have been observed in Fig.1e and the maximum translocation time has been recorded as 200 μs (indicated in the inset). The presence of spikes may be attributed to the translocation of the thrombin molecules which indicates that there is negligible non-specific adsorption. The effect of varying positive voltage has been studied after aptamer functionalization. Figs. 1(f-h) represents the current change for a concentration of 75 pM thrombin for the different voltages of +150 mV, +250 mV and +350 mV. It is observed that for both +150 and +350 mV, the change in current step is lower than that of +250 mV. This may be due to the fact that at higher voltages, a larger proportion of thrombin

translocates without being captured by the aptamer and at lower voltages, there is reduced electrophoretic force. Each of the measurements has been performed with ten nanopores.

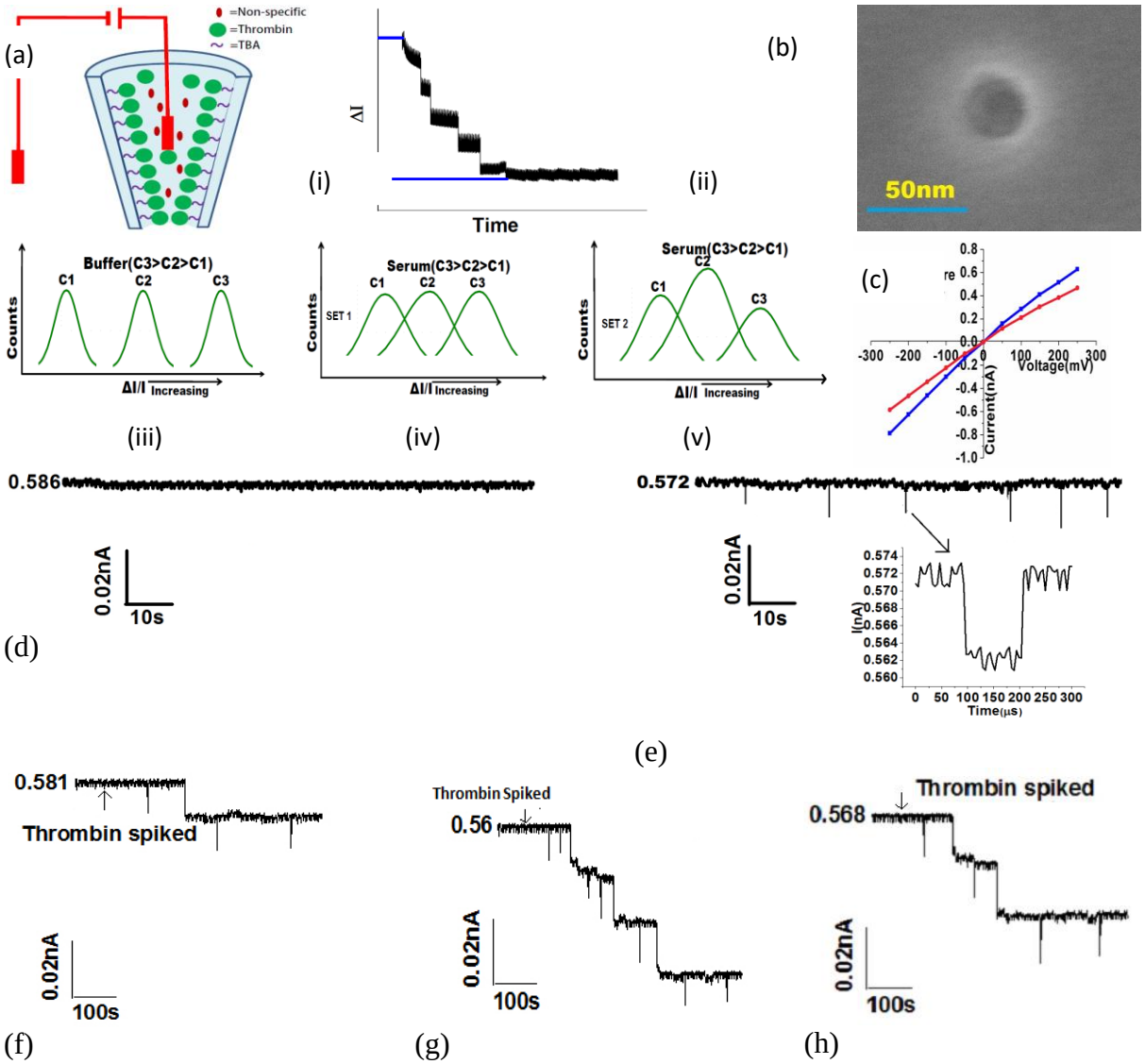

Fig. 1.1. (a)The schematic representation of (i) Conical nanopore device, (ii)Transient translocation time graph ,(iii-v)Histogram plot of sensitivity in buffer and serum, (b) Typical SEM image (c)conductance measurements before and after functionalization, (d) The current trace recorded with -250 mV after adding thrombin at a concentration of 50 pM (e) In presence of + 250mV, after adding thrombin at a concentration of 50 pM (f) Current change for a concentration of 75 pM thrombin for voltage of +150 mV (g) Current change for a concentration of 75 pM thrombin for voltage of +250 mV (h) Current change for a concentration of 75 pM thrombin for voltage of + 350 mV.

1.2 Nanoporous Graphene FET Sensor:

Detection of biomolecules electronically by conductance, impedance and ramperometric measurements have been extensively investigated. The various nanostructures fabricated for the purpose includes nanowires, nanoribbons, ordered and random nanopores, nanotubes, nanorods and nanoparticles. The substrates commonly deployed for realizing these nanostructures include silicon and its oxide, indium oxide, graphene,carbon nanotubes, anodized alumina, zinc oxide and various polymers. Nanostructured substrate not only improves the binding efficiency of the analyte due to large surface area to volume ratio but also shows a possible catalytic effect in the diffusion of the analyte molecules within the nanopores. In this aspect, the 3D configurations of the nanostructures are expected to be significantly more sensitive than their 2D counterparts, owing to larger surface area to volume ratio and increased diffusion flux. Graphene being an intrinsically zero band gap material, its band gap can be tailored significantly by applying external electric field or by surface adsorption of charged biomolecules, resulting in a significant modulation of the transconductance. This feature enables superior performance with FET configuration. Thus, in particular, 3D graphene has been recently explored in the form of nanorods, nanowalls and nanogrids for ultrasensitive detection of various biomolecules like proteins, DNA, virus and others. For example, it has been reported that reduced graphene nanowalls, fabricated by electrophoretic deposition on graphite electrode from a suspension containing Mg^{2+}-graphene oxide nanosheets results in ultra-high resolution electrochemical biosensor for detection of four bases of DNA using differential pulse voltammetry. This has been possible due to increased edge defects and high surface area to volume ratio. Further, the authors have fabricated smooth 3D graphene nanogrids by a reliable and scalable chemical method which has resulted in attomolar sensitivity in a FET configuration. The improved sensitivity was attributed to the combined effects of insignificant line edges, quantum dot like transport behaviors and improved interaction of the biomolecules within the nanopores due to the concave curvature. However, there are only few reports of investigating graphene FET biosensors in physiological solutions.

1.2.1 Working:

For sensing in physiological solutions, one of the major challenges is overcoming the Debye screening effect. This problem has been addressed by two methods: applying a high frequency signal between the source and drain which results in the breakdown of the electric-double-layer (EDL) near the surface of the FET channel, due to fast switching of the direction of the applied bias, leading to deeper penetration of the electric potential of the target protein and another technique is coating the channel surface with polyethylene glycol which increases the effective screening length. However, the intrinsic transconductance decreases in the latter due to increased thickness of the immobilized layer. These techniques have been applied in graphene field effect transistors. Gao and colleagues have reported sensing of prostate specific antigen in physiological solutions using graphene FET biosensors but the detection limit is in nM range. Graphene sensor has been used for detection of Zika virus, protein biomarkers and lead ions in serum but the detection limit is either in nanomolars or

few hundreds of femtomolars. The low sensitivity may be attributed to the 2D nature of the graphene substrate. It is expected that 3D graphene nanostructures will enable few femomolar sensing of biomolecules in serum. But for real time quantification of biomolecules in serum down to few femtomolars, the problem may be non-trivial. There is a possibility of overlap of the drain-source current sensitivity values between different concentration of the target biomolecule, especially near the lower range since in presence of serum containing large concentration of non-target molecule, there may be significant deviations in the number of target antigen captured and hence the sensitivity for a fixed concentration of the target. Again, this degree of overlap can vary from one set to another set of sensor depending on the statistical variations in the sensor fabrication process. To address this multi-level uncertainty, it will be necessary to process the output by signal processing schemes incorporating probabilistic features.

1.2.2 Specification of Sensor used:

Graphene nanogrid structure has been fabricated by following two main steps: firstly nanoporous silicon oxide (NPSO) substrate has been developed by anodic etching of p-type <100> silicon wafers of 10–20 Ω cm resistivity in a double pond electrochemical bath for 30 min under a constant current source with an electrolyte mixture of hydrofluoric acid (HF) (48 wt%) and dimethyl sulfoxide (DMSO) in the ratio of 1:9 by volume. The structure has been thermally oxidized using a dry-wet-dry sequence in an oxidation furnace for1 h at 900^{o}C to obtain pores with 30 nm diameter (*d)* and 100 nm length (*l).* Secondly, grapheme has been deposited on nanoporous silicon oxide substrate by electrophoretic deposition (EPD) method. Interdigited metal electrodes have been fabricated on the substrate for EPD of graphene with aluminium by screen printing method which has been next cured at 750^{o}C for 1 min. The width of the electrodes is 60 μm and spacing between two consecutive fingers of each electrode is 100μm. This has been followed by evaporation of gold metal which improves the longevity of the sensor. The electrodes have been connected with function generator and the prepared colloidal graphene solution has been pipetted onto the substrate so that the droplet of the solution should be covering the electrodes. Then a sinusoidal voltage (peak to peak 19.5 V) has been applied for 60 s. The graphene deposited substrate has been heated at 75 °C for 2 min for increasing the adhesion between graphene and the substrate.

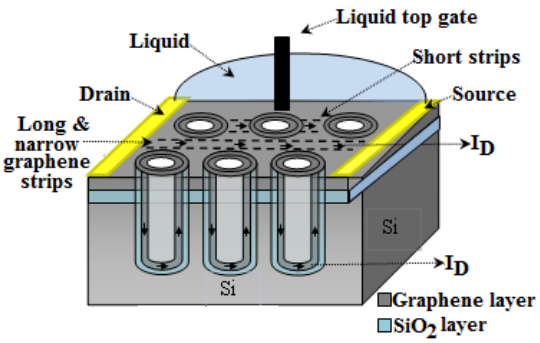

Fig.1.2a

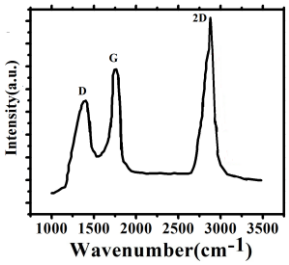

Fig.1.2b

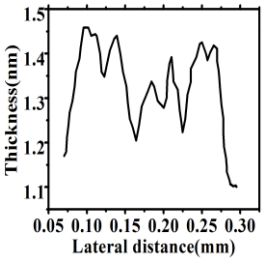

Fig.1.2c

The aforesaid fabrication results in a nanogrid structure with long and narrow graphene strips alternating with series combination of short and narrow strips on planar regions and within pores, the schematic of which is shown in Fig.1.2a. The Raman spectroscopy and surface profilometry of graphene is indicated in Fig.1.2b and 1.2c. It has been observed from Fig.1.2b that the G peak is higher than D peak which indicates low defect density. The presence of prominent 2D peak in Fig.1.2c points towards the bilayer nature of graphene.

For immobilization of anti Hep-B monoclonal antibody, (procured from Sigma Aldrich, St. Louis, MO, USA) the graphene nanogrid structure has been treated with 25% glutaraldehyde aqueous solutions (procured from Sigma Aldrich) for 4 h. Glutaraldehyde acts as a crosslinker, which helps the antibody bind covalently with the substrate by activating the negatively charged carboxylic groups of graphene. After treatment, the sensor has been rinsed thoroughly with de-ionized water for three times. Then the sensor has been incubated with antibody for 1 h followed by washing with phosphate buffer saline (PBS). To enclose the active region of the sensor, a PDMS well of approximately 2.5 mm thickness has been positioned in the region between the electrodes by observing under an optical microscope (Leica DM 2500). The width of the PDMS sheet is around 4 mm on all the sides. The PDMS sheets made from SYLGARD 184A and SYLGARD 184 B has been obtained from Department of Chemical Engineering, IIT Kharagpur. For proper sealing of the PDMS with silicon oxide substrate, oxygen plasma treatment has been done. Polyethylene tubing has been attached as the inlet and outlet of the PDMS wells through which proteins and buffer solutions have been drawn using a syringe pump. The protocol for antibody immobilization has been optimized in previous reports.

Each sensor chip has been interfaced with the PC through a sensor holder set up. In the setup, the bond pads of the sensors have been placed below the probes of the sensor holder. The probes of the sensor holder which are in contact with the drain and source electrodes are spring loaded and adjustment screws have been provided to align them with the bond pads. The gate electrode has been inserted through a small hole in the setup and soldered to a metal connector. The picture of the setup is shown in Fig.1.3.

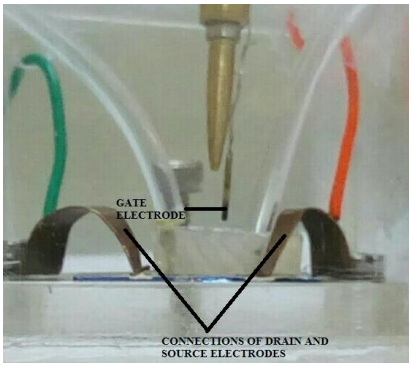

Fig.1.3

The gate electrode has been connected to a dc power supply (Keysight technologies). A high frequency signal of 20 mV amplitude in the range of 200 kHz to 1000 kHz has been applied through a signal generator (Agilent 33521A).It has been modulated by a reference signal of 1.43 kHz using a lock-in amplifier(Stanford Research, SR 830) and fed to the graphene FET at the source. The root mean square value of the mixing current (I_{mix}) through the drain terminal has been detected at the modulated frequency for a total duration of around 30 minutes, using the PC interface of the lock-in amplifier. both before and after application of the analyte. Online computational analysis has been carried out in the same PC with MATLAB software for data analysis. The total analysis process has been divided in two phases: the calibration phase and testing phase. In the calibration phase, sensitivity of $I_{mix}(S)$ has been evaluated for various concentrations of the target antigen and then processed. In the testing phase, the unknown sensitivity value has been correlated with the concentration of the target antigen.

Hep- B surface antigen has been used as the target molecule which has been procured from US Biological, USA (1 mg/ml).For the purpose of measurement, serial dilutions of Hep-B antigen solution has been performed in PBS with pH 7.2 and 100 mM ionic strength to produce concentrations ranging from 0.1fM to 1 pM of Hep-B. To execute calibration and testing, separate buffer solution of these antigens have been measured. For measurements in physiological analyte, Hep-B surface antigen has been spiked in blood serum. The blood samples collected from a microbiological laboratory have been kept for some time at room temperature to separate the serum. The concentration of Hep-B has been limited to 1 pM, since the current output saturates for a higher concentration. Each of the samples has been tested in five batches. In each batch, ten sensors have been use. Mean and standard deviations have been recorded, both inter and intra batches.

Chapter 2

Artificial Neural Networks

The human brain is a highly complicated machine capable of solving very complex problems. Although we have a good understanding of some of the basic operations that drive the brain, we are still far from understanding everything there is to know about the brain.

In order to understand Artificial Neural Networks (hereby referred to as ANN), it is required to have a basic knowledge of how the internals of the brain work. The brain is part of the central nervous system and consists of a very large Neural Network (hereby referred to as NN). The NN is highly complicated, of which a very simple model is discussed which is required for understanding ANN.

2.1 The Human Neuron:

The centre of the neuron is called the nucleus. The nucleus is connected to other nucleuses by means of the dendrites and the axon. This connection is called a synaptic connection. The neuron can fire electric pulses through its synaptic connections, which is received at the dendrites of other neurons.

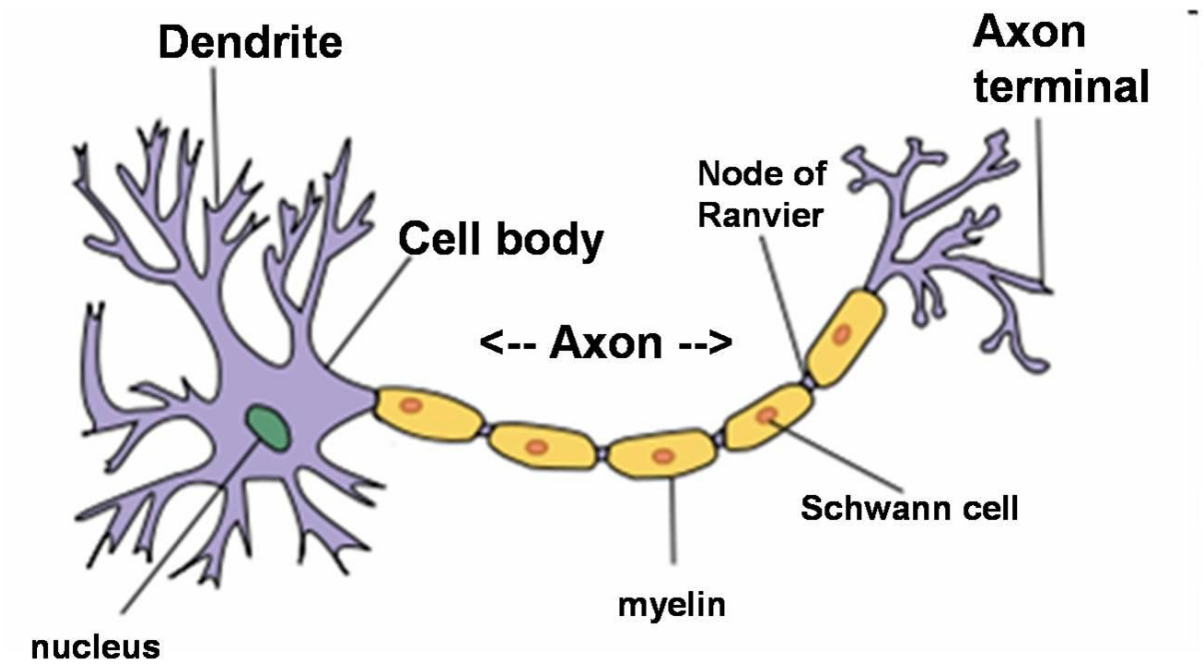

Fig. 2.1 Biological Neuron

When a neuron receives enough electric pulses through its dendrites, it activates and fires a pulse through its axon, which is then received by other neurons. In this way information can

propagate through the NN. The synaptic connections change throughout the lifetime of a neuron and the amount of incoming pulses needed to activate a neuron (the threshold) also change. This behaviour allows the NN to learn.

The human brain consists of around 10^{11} neurons which are highly interconnected with around 10^{15} connections [2]. These neurons activates in parallel as an effect to internal and external sources.

2.2 Artificial Neuron:

A single artificial neuron can be implemented in many different ways. The general mathematic definition [3] is as shown.

$$y(x) = f\left(\sum_{i=0}^{n} w_i x_i\right)$$

Where x is a neuron with n input dendrites x_0, x_1, ..., x_n and one output axon $y(x)$ where w_0, w_1, ..., w_n are weights determining how much the inputs should be weighted. Here $f(.)$ is the activation function.

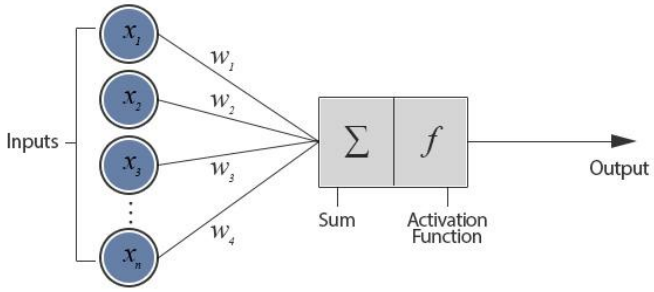

Fig. 2.2 Artificial Neuron

The activation function determines how strong the output of the neuron should be. Generally, an output of 1 is considered to be strongest and 0 to be weakest. That is, an activation function maps the input between 0 and 1, such as sigmoid function. There are also other activation functions, for which this limit is [-1, 1] etc., such as hyperbolic tangent function (Each necessary activation function used in this thesis is explained in their respective sections).

2.3 Artificial Neural Network:

An ANN is an interconnection between many artificial neurons. There are a huge number of ANN models which can deal with different types of problems. These networks are trained with their respective datasets to predict output for data which is of similar type as that of the training dataset but not necessarily a part of it. In this thesis different types of ANN are used and are discussed in their respective sections.

Chapter 3

Fuzzy Logic

Fuzzy logic [4] starts with and builds on a set of user-supplied human language rules. The fuzzy systems convert these rules to their mathematical equivalents. This simplifies the job of the system designer and the computer, and results in much more accurate representations of the way systems behave in the real world.

Fuzzy logic models, called fuzzy inference systems, consist of a number of conditional "if-then" rules. For the designer who understands the system, these rules are easy to write, and as many rules as necessary can be supplied to describe the system adequately (although typically only a moderate number of rules are needed).

In fuzzy logic, unlike standard conditional logic, the truth of any statement is a matter of degree. (How cold is it? How high should we set the heat?) We are familiar with inference rules of the form p => q (p implies q). With fuzzy logic, it's possible to say (.5* p) => (.5 * q). For example, for the rule: *if (weather is cold) then (heat is on)*, both variables, *cold* and *on* map to ranges of values. Fuzzy inference systems rely on membership functions to explain to the computer how to calculate the correct value between 0 and 1. The degree to which any fuzzy statement is true is denoted by a value between 0 and 1.

3.1 Fuzzy Sets:

Fuzzy Set Theory was formalised by Professor Lofti Zadeh at the University of California in 1965. What Zadeh proposed is very much a paradigm shift that first gained acceptance in the Far East and its successful application has ensured its adoption around the world.

Bivalent Set Theory can be somewhat limiting if we wish to describe a 'humanistic' problem mathematically. For example, Fig. 2.1 below illustrates bivalent sets to characterise the temperature of a room.

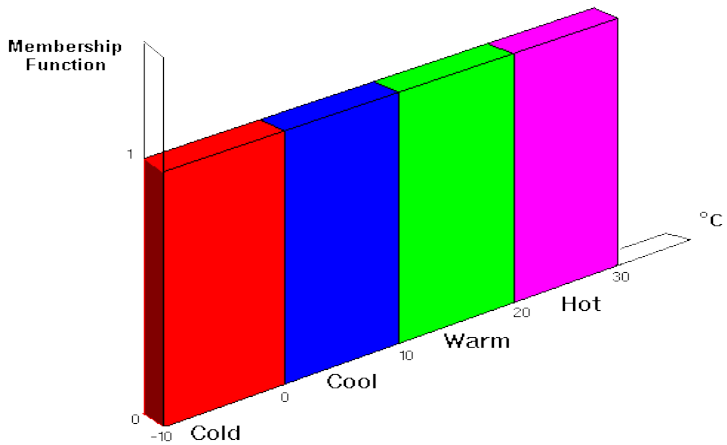

Fig. 3.1 Bivalent sets to characterise the temperature of the room

The most obvious limiting feature of bivalent sets that can be seen clearly from the diagram is that they are mutually exclusive - it is not possible to have membership of more than one set. Opinion would widely vary as to whether 0 degrees Centigrade is 'cold' or 'cool' hence the expert knowledge we need to define our system is mathematically at odds with the humanistic world. Clearly, it is not accurate to define a transition from a quantity such as 'warm' to 'hot' by increasing 1 degree Centigrade. In the real world a smooth drift from warm to hot would occur.

This natural phenomenon can be described more accurately by Fuzzy Set Theory. Fig.2.2 below shows how fuzzy sets quantifying the same information can describe this natural drift.

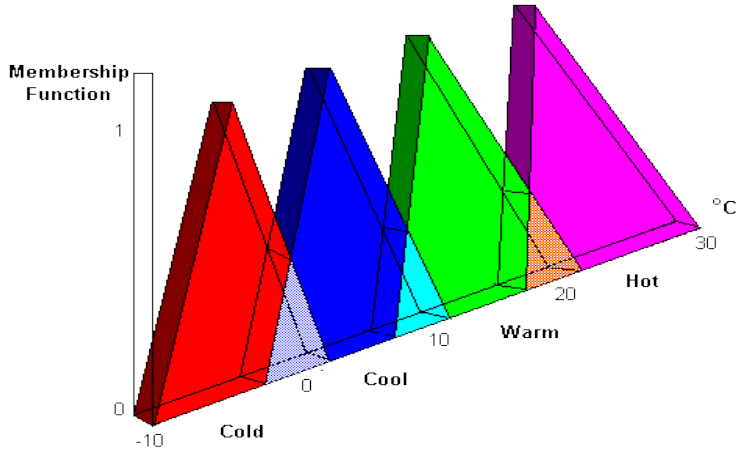

Fig. 3.2 Fuzzy sets to characterise the temperature of the room

3.2 Membership Functions:

As shown in Fig. 2.2 each of the triangles represents a membership function. The membership values of an input for any set can be found by the y-intersect of the functions for a certain input. That is,

$$y = \mu(x)$$

Where $\mu(.)$ is the membership function and x is the input.

Chapter 4

Quantification of Biomolecules in Complex Analyte in Nanopores: Probabilistic Fuzzy system

Signal processing tools like fuzzy logic, neural networks, progressive least square fitting and others have been applied by various researchers in chemical and biological sensors in practical settings but these sensors either involved washing or recovery steps or were not real time based operation, which reduced the uncertainties in the electrical outputs. However, probabilistic fuzzy logic has been applied by other researchers to model stochastic biological phenomena with uncertain kinetic parameters and for cancer network analysis [5-7].

In this study, probabilistic fuzzy model based on Monte Carlo simulation (PF-MC) is introduced to quantify the presence of thrombin in serum down to 50 pM concentrations using aptamer coated glass nanopores. The overlap between 50 pM and 75 pM concentration has been observed to be as high as 78%. The system has been tested with twenty five spiked serum samples using five nanopores each. The model has been satisfactorily observed to classify the target protein in the correct range.

The sensitivity value of the sensor changes (hereby referred to as drift) with consecutive data acquisition, even for the same concentration. This drift in the sensitivity results in high overlap among the sensitivities recorded for other concentrations. This problem can be addressed by using ***Fuzzy Logic*** based systems. But the amount of drift and even the mean of the distribution of the sensitivity values changes for the same setup with same target concentration at different instances of time. This change in drift is probabilistic in nature and follows a Gaussian distribution. Due to this probabilistic nature, simple Fuzzy Logic based approach would not be appropriate and to address this issue, ***Probabilistic Fuzzy Logic*** based system is used.

In this approach the probability is incorporated in the membership function of the fuzzy system itself as explained in the following sections.

4.1 Membership Function:

A membership function is generated for each of the recorded concentration values. Since the sensitivity value is dependent upon the concentration, thus the normalised and fitted histogram of the sensitivity values for a certain concentration can be taken as the membership function for that target concentration.

These types of drift usually follow a Gaussian or Normal distribution. Although the Gaussian nature is not apparent in the raw histogram (i.e. treating each sensitivity value as unique) [Fig

4.1(a)] due to small number of data, when the values are clustered together with a certain number of bins (3 or 4 in this case), the Gaussian nature becomes apparent [Fig. 4.1(b)]. In the figure below, X axis shows the sensitivity values and Y axis is the frequency of occurrence.

These Gaussian functions vary in amplitude since they are fitted from histograms [Fig. 4.2]. And thus are not suitable to be turned into membership functions unaltered. For this, the functions are linearly normalised individually with respect to the maximum value of the function by mapping it to unity [Fig. 4.3].

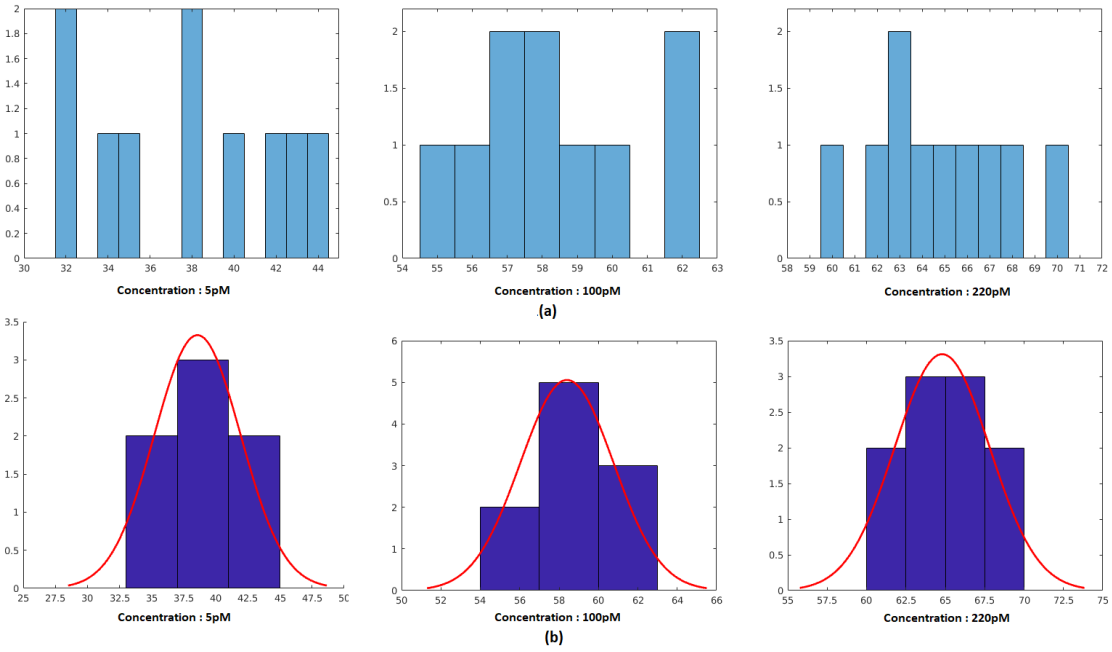

Fig. 4.1

The probability is incorporated in these membership functions. The mean (μ_i) and standard deviation (σ_i) is calculated from the experimentally obtained histogram after fitting them to a Gaussian function. Since the mean (μ_i) and standard deviation (σ_i) has a drift of around 5-10% as mentioned before, thus for every iteration, μ_i and σ_i of the Gaussian membership function is changed using random values from two more Gaussian functions (one for randomising the μ_i and one for the σ_i).

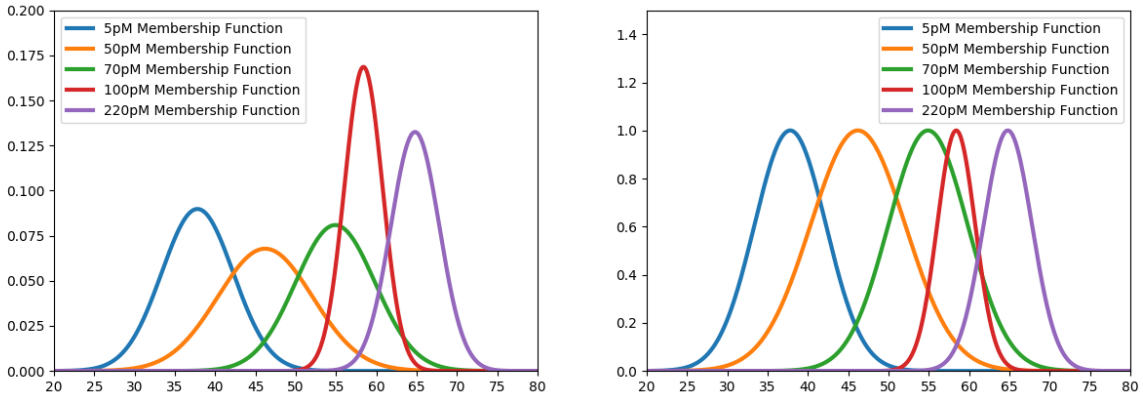

Fig. 4.2 and 4.3

The Gaussian function for randomising mean and standard deviation take μ_i and σ_i as means respectively and a 5% standard deviation (σ_d) to generate the randomised mean ($\mu_i{}'$) around μ_i and randomised standard deviation ($\sigma_i{}'$) around σ_i.

4.2 Monte-Carlo Process:

The $\mu_i{}'$ and $\sigma_i{}'$ are generated several times to take into consideration a number of possible $\mu_i{}'$ and $\sigma_i{}'$ and choose an average value for both ($\overline{\mu_i}{}'$ & $\overline{\sigma_i}{}'$). The number of times the process is repeated is bounded by 2 factors. The number of repetition (r) should be such that,

$$\text{For } r_{max} \begin{cases} \left|\overline{\mu_i{}'}\right| \geq \mu_i{}' \pm \frac{\sigma_d}{5} \\ \left|\overline{\sigma_i{}'}\right| \geq \sigma_i{}' \pm \frac{\sigma_d}{5} \end{cases}$$

$$\text{For } r_{min} \begin{cases} \left|\overline{\mu_i{}'}\right| \leq \mu_i{}' \pm 3\sigma_d \\ \left|\overline{\sigma_i{}'}\right| \leq \sigma_i{}' \pm 3\sigma_d \end{cases}$$

It was experimentally determined that $r \in [20,30]$ (where, r is an integer), follows the above relations acceptably and was considered to be 25 to determine all the results.

4.3 De-fuzzification:

Since the output (concentration) is not present in the membership function, i.e. the output is not a fuzzy function itself, but is completely depended upon the fuzzy variables, thus standard defuzzification algorithms cannot be used to get a concentration value. Thus to get a singular concentration value from the fuzzy membership value, the following formula is used,

$$\bar{c} = \frac{\sum c_i m_i}{\sum m_i}$$

Where, c_i = i^{th} concentration; m_i = membership value from membership function corresponding to i^{th} concentration.

4.4 Algorithm:

The algorithm can be summarized as follows,

1) Take a single sensitivity value as input.
2) Perform Monte-Carlo on μ_i and σ_i to obtain $\left|\overline{\mu_\iota'}\right|$ and $\left|\overline{\sigma_\iota'}\right|$.
3) Generate Gaussian functions with $\left|\overline{\mu_\iota'}\right|$ and $\left|\overline{\sigma_\iota'}\right|$ and normalise to obtain probabilistic fuzzy membership functions.
4) Calculate membership values for the input from membership functions obtained from the previous step.
5) De-fuzzify using de-fuzzification method described above.

4.5 Results:

This method is particularly helpful in measuring the membership values at the intersect points of two or more membership functions, where membership values for those membership functions are same and no decision can be taken while using standard fuzzy logic.

To demonstrate this, values at and around dominating intersect points (i.e. those intersects where the value at the intersection point is higher than membership values of other concentrations) of a non-probabilistic model were taken and are compared with the probabilistic version (for 2 separate iterations) and the results are shown in Table 4.1.

[In the table membership values of interest are shown in bold]

Table 4.1

Sensitivity	**Non-Probabilistic Model Membership Values**	**Probabilistic Model Membership Values [Iteration 1]**	**Probabilistic Model Membership Values [Iteration 2]**
41.314	**5pM = 0.731** **50pM = 0.708** 70pM = 0.022 100pM = 4.782e-12 220pM = 6.157e-14	**5pM = 0.762** **50pM = 0.701** 70pM = 0.030 100pM = 3.859e-12 220pM = 8.589e-14	**5pM = 0.748** **50pM = 0.715** 70pM = 0.023 100pM = 1.105e-12 220pM = 2.019e-14
41.414	**5pM = 0.718** **50pM = 0.718** 70pM = 0.023 100pM = 6.483e-12 220pM = 7.973e-14	**5pM = 0.785** **50pM = 0.713** 70pM = 0.025 100pM = 6.360e-12 220pM = 6.436e-14	**5pM = 0.725** **50pM = 0.702** 70pM = 0.027 100pM = 1.095e-11 220pM = 8.387e-15
41.514	**5pM = 0.705** **50pM = 0.728** 70pM = 0.025 100pM = 8.772e-12 220pM = 1.031e-13	**5pM = 0.677** **50pM = 0.714** 70pM = 0.025 100pM = 4.493e-12 220pM = 3.400e-14	**5pM = 0.697** **50pM = 0.734** 70pM = 0.027 100pM = 1.646e-12 220pM = 1.843e-13

50.8328	5pM = 0.013 **50pM = 0.733** **70pM = 0.712** 100pM = 0.006 220pM = 2.127e-05	5pM = 0.0117 **50pM = 0.696** **70pM = 0.741** 100pM = 0.013 220pM = 2.195e-05	5pM = 0.013 **50pM = 0.700** **70pM = 0.765** 100pM = 0.004 220pM = 1.898e-05
50.9328	5pM = 0.012 **50pM = 0.723** **70pM = 0.723** 100pM = 0.006 220pM = 2.479e-05	5pM = 0.017 **50pM = 0.717** **70pM = 0.704** 100pM = 0.005 220pM = 3.029e-05	5pM = 0.012 **50pM = 0.714** **70pM = 0.684** 100pM = 0.007 220pM = 1.121e-05
51.0328	5pM = 0.011 **50pM = 0.713** **70pM = 0.735** 100pM = 0.007 220pM = 2.888e-05	5pM = 0.014 **50pM = 0.716** **70pM = 0.735** 100pM = 0.004 220pM = 4.298e-05	5pM = 0.009 **50pM = 0.683** **70pM = 0.732** 100pM = 0.005 220pM = 3.665e-05
57.1656	5pM = 7.466e-05 50pM = 0.176 **70pM = 0.899** **100pM = 0.872** 220pM = 0.040	5pM = 4.840e-05 50pM = 0.170 **70pM = 0.869** **100pM = 0.852** 220pM = 0.039	5pM = 7.611e-05 50pM = 0.164 **70pM = 0.873** **100pM = 0.846** 220pM = 0.039
57.2656	5pM = 6.767e-05 50pM = 0.170 **70pM = 0.891** **100pM = 0.891** 220pM = 0.043	5pM = 5.701e-05 50pM = 0.169 **70pM = 0.916** **100pM = 0.951** 220pM = 0.032	5pM = 5.191e-05 50pM = 0.159 **70pM = 0.898** **100pM = 0.927** 220pM = 0.046
57.3656	5pM = 6.129e-05 50pM = 0.165 **70pM = 0.882** **100pM = 0.908** 220pM = 0.047	5pM = 7.440e-05 50pM = 0.178 **70pM = 0.848** **100pM = 0.863** 220pM = 0.041	5pM = 6.232e-05 50pM = 0.196 **70pM = 0.888** **100pM = 0.937** 220pM = 0.047
61.1164	5pM = 1.041e-06 50pM = 0.040 70pM = 0.451 **100pM = 0.517** **220pM = 0.473**	5pM = 1.493e-06 50pM = 0.032 70pM = 0.478 **100pM = 0.525** **220pM = 0.438**	5pM = 1.162e-06 50pM = 0.038 70pM = 0.471 **100pM = 0.544** **220pM = 0.513**
61.2164	5pM = 9.249e-07 50pM = 0.038 70pM = 0.440 **100pM = 0.492** **220pM = 0.492**	5pM = 9.399e-07 50pM = 0.040 70pM = 0.441 **100pM = 0.548** **220pM = 0.442**	5pM = 1.010e-06 50pM = 0.033 70pM = 0.439 **100pM = 0.531** **220pM = 0.519**
61.3164	5pM = 8.212e-07 50pM = 0.036 70pM = 0.428 **100pM = 0.468** **220pM = 0.512**	5pM = 5.569e-07 50pM = 0.036 70pM = 0.407 **100pM = 0.388** **220pM = 0.505**	5pM = 7.099e-07 50pM = 0.046 70pM = 0.505 **100pM = 0.541** **220pM = 0.474**

As it can be seen from the table that the sensitivity values (e.g. 41.414, 50.9328 etc.) where standard fuzzy logic failed, this model was able to give out a proper decision in such scenarios.

Chapter 5

Biomolecule Detection in Complex Analyte using Nanoporous Graphene FET Sensor: Probabilistic Neural Networks

For sensitivity values ranging from 10.8% to 75%, the concentration varies from 0.1fM to 1000fM; with majority of the data residing in lower concentration range. So, any standard curve fitting model will try to minimize the overall error thus getting biased towards higher concentration range. In turns giving better accuracy at higher concentration and significantly sacrificing accuracy in lower concentration (this effect is further discussed later).

For the acquired data, there is an overlap of around 30%. This prohibits the use of standard non-probabilistic neural network (NN) models, since the accuracy will suffer greatly. Also, standard NN models will not be able to take into account the inherent drift of the sensor. To circumvent these problems a Probabilistic NN model is proposed in this paper.

The whole data analysis is divided into several steps as discussed below.

5.1 Probabilistic activation function:

There are different methods of creating a probabilistic neural network. The probability can be incorporated in the data itself, while the NN remains static [10]. The probability can be incorporated in the input weight vector of one or more neurons and keeping the data static [9]. In this paper we explore a model in which the activation functions of hidden layer neurons are made probabilistic, thus making the whole network dynamic, while the data and weights remain static.

In this paper we use probabilistic tan-hyperbolic activation function to achieve the intended goal.

Standard tan-hyperbolic activation function [Fig.5.1a] is defined as

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

In case of probabilistic tan-hyperbolic, the equation becomes as follows

$$g(x) = f(p.x) = \frac{e^{px} - e^{-px}}{e^{px} + e^{-px}}$$

Where, *p* is a random variable from a chosen probability distribution. For example, if *p* is chosen from a normal distribution with mean = 1 and standard deviation = 10%, then over a number of iterations [1000 in this example] the probabilistic nature can be observed as shown in Fig.5.1b In this figure, the transparency is inversely proportional to the probability, i.e. higher the transparency, lower the probability of that particular curve appearing.

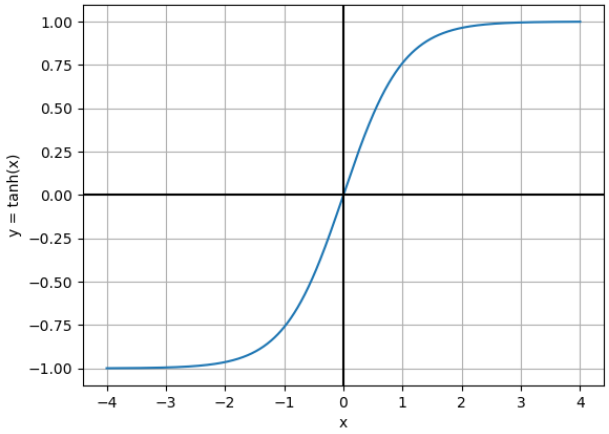

Fig.5.1a

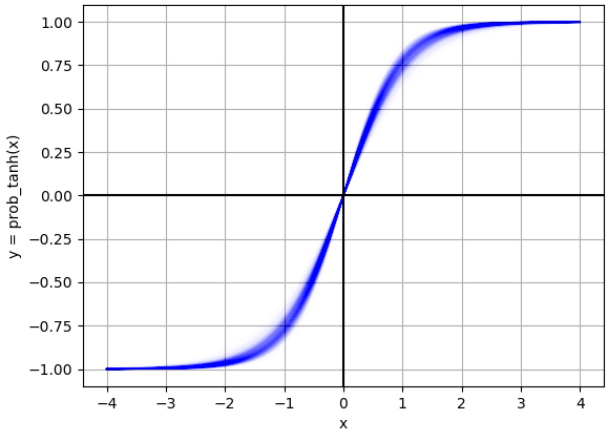

Fig.5.1b

5.2 Scale Correction:

As it can be seen from Fig. 5.2a, the concentration follows a near exponential relation with sensitivity. Given the range of 0.1-1000, it is safe to assume the base of the exponent to be 10. Since many concentration values are <1, so, taking log of the concentration will result in negative values, which is not suitable for NN training and testing purposes. Thus we need to push up the lowest value to zero by adding a zero-correction factor. The resultant curve is comparatively linear and resides in the first quadrant as shown in Fig.5.2b. The scale correction makes the variations of sensitivity with concentration more apparent in the lower

concentration range, which was not the case for the raw data. The data is considered to be piece-wise linear in the ranges given in Table-5.1.

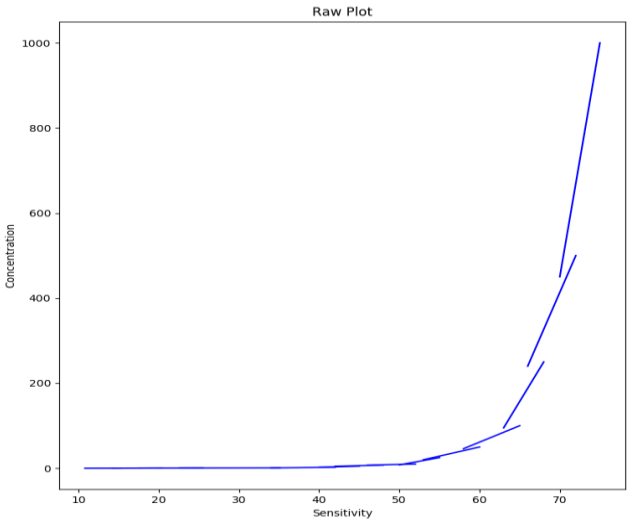

Fig.5.2a

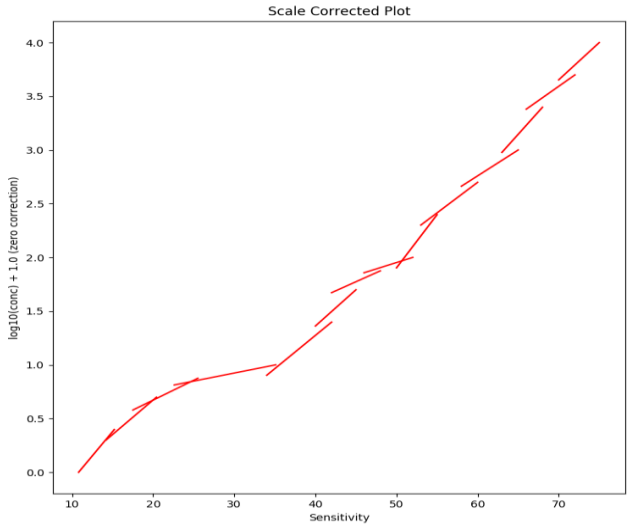

Fig.5.2b

5.3 Classification:

Each linear line segment as shown in Fig.5.2 is considered to be a class. It is evident that there is a significant amount of overlap (~30%) between adjacent classes (when a single sensitivity value corresponds to more than one concentration, then it is considered to be in one of the overlapped regions). Thus, with a single input parameter (sensitivity) it is not possible to classify the data with an accuracy of more than ~70% with static machine learning (ML) or neural network models. This drawback of static ML and neural net is overcome by implementing probabilistic neural network classifier.

The data is classified in 14 classes based on sensitivity values, each line segment representing a class. The output is converted into a Dirac-Delta vector [20] of 14 variables with corresponding class element for a certain class equal to 1 and the rest equal to 0. For example, if an input belongs to class 2 [indexing from 0], then the Dirac-Delta output vector for that input point would be as follows: [0 0 1 0 0 0 0 0 0 0 0 0 0 0].

Next a classifier neural network, consisting of 3 hidden layers with 14 neurons in each layer with probabilistic tan-hyperbolic activation function and an output layer with 14 neurons (for 14 classes) and *softmax* activation function is trained with the data with *categorical-crossentropy* as the loss function and *Adam* optimizer. The softmax function is required at the output stage so that the modulus of the Dirac-Delta output vector remains 1.

In this model, the accuracy of a single iteration is ~70%, similar to a static model. But this being a probabilistic model, the output is to be generated over a number of iterations. This significantly increases the accuracy to >99%. As it can be seen from Fig.5.3, the model was able to classify regions with significant overlap with an appreciable accuracy by probabilistically choosing the class for an input in the overlapped region over 50 iterations.

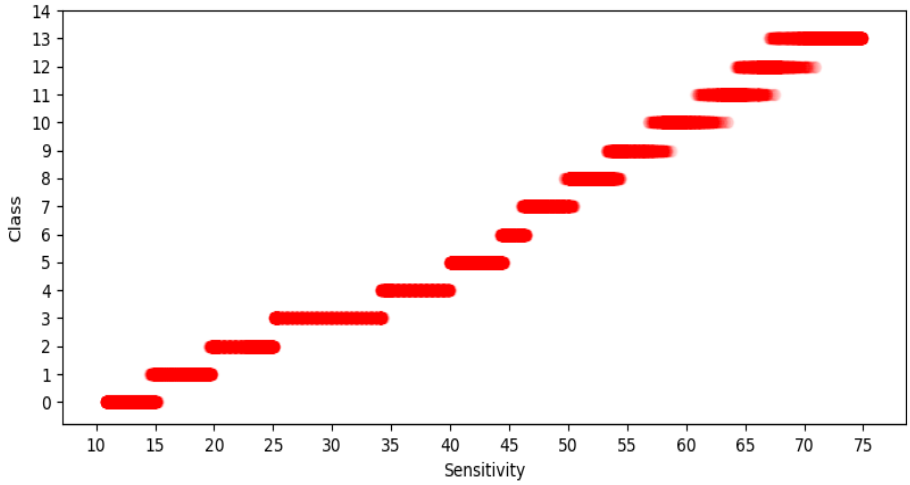

Fig. 5.3

5.4 Concentration Estimation:

This is the final neural network block, comprising of 14 disjoint neural networks for 14 classes. These take the sensitivity value (normalized to 1) and the class (output of the classifier) to predict the concentration value in the log scale (section b.), which is then used to generate the actual concentration value.

Each network in this block is dedicated to a certain class and comprises of 2 hidden layers with 4 neurons in each running a probabilistic tan-hyperbolic function. These are single input

single output networks and the output of the classifier is simply used to select the appropriate network. Since these are also probabilistic networks, thus for these also the output is generated over a certain number of iterations. Fig.5.4 shows the relation between input and output for the whole block with the classes entered manually from training data for 50 iterations with the red lines showing the original plot of training data and green showing the output of the probabilistic block with the transparency inversely proportional to the probability.

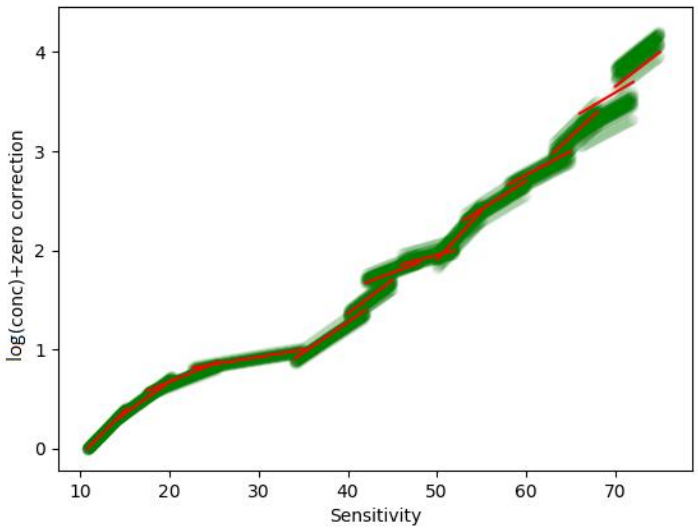

Fig.5.4

5.5 Algorithm:

We can summarize the complete algorithm as follows:

Step 1: Take sensitivity value as input (*x*)

Step 2: Select number of iteration (*n*)

Step 3: Pass *x* to classifier and estimate the class

Step 4: Use the class to select the particular estimator NN

Step 5: Pass the input to the chosen NN and get concentration

Step 6: Repeat Step 3 to 5, *n* number of times to get *n* outputs

Step 7: Correct the scale

Step 8: Take the mean of these *n* outputs to get the mean output and take the minimum and maximum of the values to get the range of concentration for the input sensitivity; which takes

into account the overlap in data through the classifier block and also the inherent drift of the sensor through the estimator block.

5.6 Results:

Table 5.2 shows the final results obtained from different estimation method to map the sensitivity values to concentration. These methods are discussed and compared with Probabilistic NN model below.

5.6.1 Polynomial Fitting:

Fig.5.5 shows the mean-squared-error (mse) of the fit with degree of the polynomial. It was observed that the mse saturates after 6^{th} degree and increasing the degree further did not change the mse significantly. Fig.5.6 shows the 6^{th} degree polynomial fit of the data.

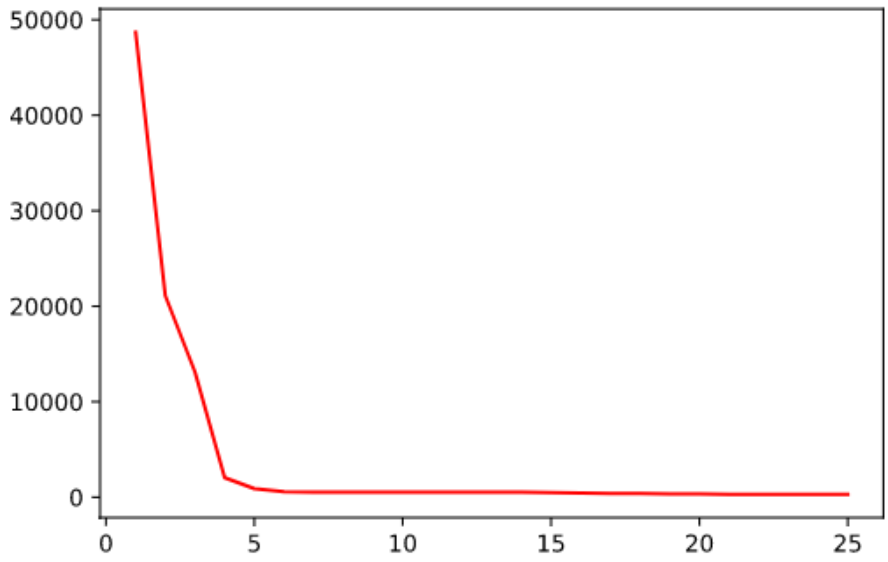

Fig. 5.5

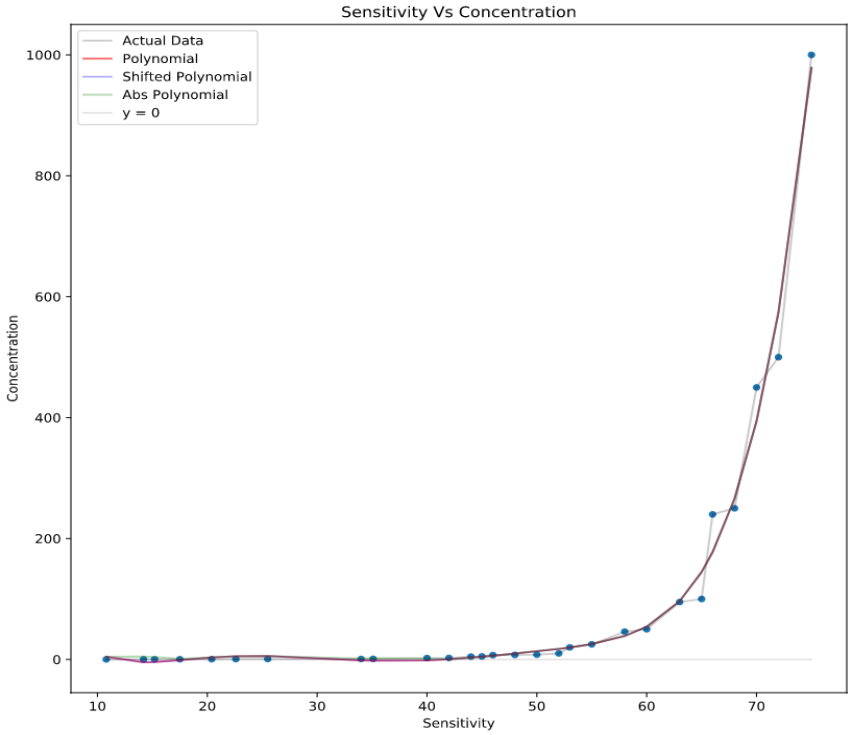

Fig. 5.6

5.6.2 Static Neural Network:

To further compare the results with a static NN, we chose a single layer network with tan-hyperbolic activation function until mse decreased to an acceptable range, but since overall mse is decreased and the lower concentration region in not prioritized, the lower concentration range suffers from a large deviation from the actual value. The fit of the NN is shown in Fig. 5.7.

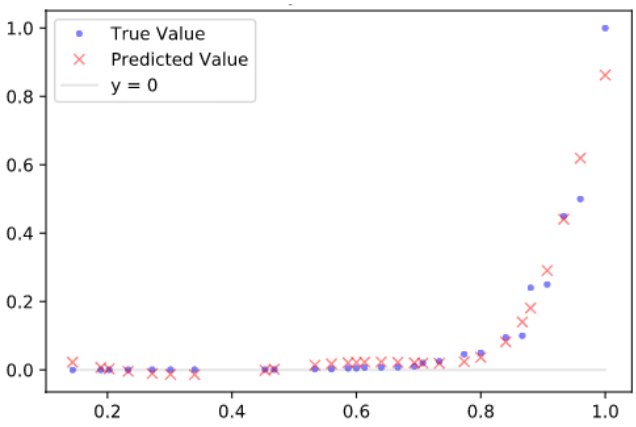

Fig.5.7 X axis: Sensitivity, Y axis: Concentration (Axes normalized to 1)

5.6.3 Probabilistic Neural Network:

Fig.5.8 shows the output of the probabilistic model proposed in this paper over 50 iterations. The spread of the green scatter plot in the y axis shows the range of output that is obtained for each sensitivity value with the transparency inversely proportional to the probability of achieving that value.

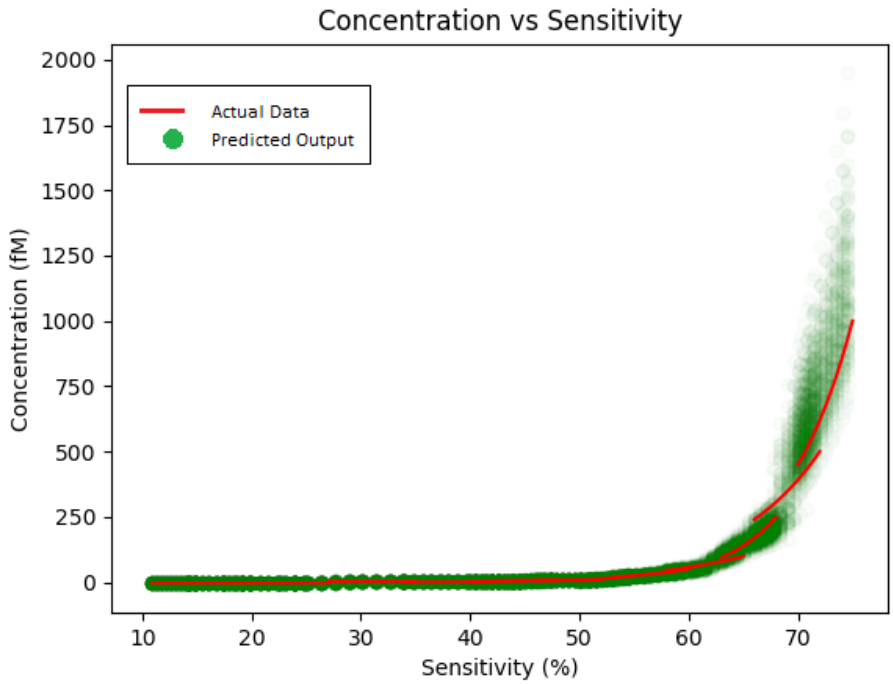

Fig. 5.8

TABLE-5.1
SENSITIVITY VALUES FOR DIFFERENT ANTIGEN CONCENTRATION

Concentration of Hep-B (fM)	Range of sensitivity (%)	Standard Deviation (%)
0.1-0.25	10.8 - 15.2	3.5
0.2-0.5	14.2-20.4	3.8
0.38-0.75	17.5-25.5	2.5
0.65-1	22.6-35.1	2.3
0.8-2.5	34-42	2.5
2.3-5	40-45	3.0
4.7-7.5	42-48	3.4
7.2-10	46-52	3.5
8-25	50-55	2.4
20-50	53-60	2.2
46-100	58-65	2.3
95-250	63-68	2.4
240-500	66-72	2.2
450-1000	70-75	2.3

TABLE-5.2
OUTPUT OF DIFFERENT ESTIMATION MODELS

Sensitivity (%)	Polynomial (Deg. 6) (fM)	Single Layer Neural Network Model (fM)	Probabilistic Neural Network Model	
			Output (fM) [Average]	Output (fM) [Range]
11	3.261	21.300	0.104	0.103 - 0.104
11.5	0.678	18.783	0.115	0.114 - 0.115
12	1.308	16.352	0.128	0.127 - 0.129
12.5	2.771	14.010	0.142	0.141 - 0.144
13	3.779	11.756	0.159	0.155 - 0.161
13.5	4.397	9.592	0.176	0.173 - 0.179
14.5	4.687	5.539	0.217	0.209 - 0.227
15	4.461	3.651	0.238	0.22 - 0.253
16	3.490	0.156	0.27	0.26 - 0.292
18	0.453	5.688	0.368	0.34 - 0.396
20	2.680	9.985	0.437	0.414 - 0.471
23	5.571	13.532	0.546	0.51 - 0.583
25	5.860	14.018	0.679	0.632 - 0.736
30	2.280	9.366	0.821	0.758 - 0.874
35	2.044	1.338	0.881	0.791 - 0.982
40	1.603	13.590	2.18	1.617 - 2.792
45	4.479	21.820	6.538	5.756 - 7.829
50	13.494	22.202	8.372	5.966 - 10.057
60	54.207	36.641	46.805	37.151 - 58.916
62	78.990	62.153	71.923	49.35 - 96.387

References

1. Bell, N. A. W.; Thacker, V. V.; Hernández-Ainsa, S.; Fuentes-Perez, M. E.; Moreno-Herrero, F.; Liedl, T.; Keyser, U. F. Multiplexed ionic current sensing with glass nanopores. Lab on a Chip 2013, 13 (10), 1859.
2. Tettamanzi, A. and Tomassini, M., *Soft Computing*, Springer-Verlag (2001).
3. Simon Haykins, *Neural Network and Learning Machines, 3ed,* Pearson (2009).
4. Timothy J. Ross, *Fuzzy Logic with Engineering Applications, 3ed,* Wiley (2010).
5. Liu, Z.; Li, H.-X. A probabilistic fuzzy logic system for modeling and control. IEEE Transactions on Fuzzy Systems 2005, 13 (6), 848–859.
6. Pathway and network analysis of cancer genomes. *Nature Methods* **2015,** 12 (7), 615-621.
7. Morris, M. K.; Saez-Rodriguez, J.; Clarke, D. C.; Sorger, P. K.; Lauffenburger, D. A. Training Signaling Pathway Maps to Biochemical Data with Constrained Fuzzy Logic: Quantitative Analysis of Liver Cell Responses to Inflammatory Stimuli. PLoS Computational Biology 2011, 7 (3).
8. E. C. Yusko , B.R.Bruhn , O.M. Eggenberger , J.Houghtaling , R.C.Rollings, N.C.Walsh , S.Nandivada , M.Pindrus , A.R.Hall , D.Sept , J.Li, D.S.Kalonia and M.Mayer, "Real-time shape approximation and fingerprinting of single proteins using a nanopore", Nature Nanotechnology, published online 2016.
9. D. F. Specht, "Probabilistic neural networks," *Neural Networks*, vol. 3, no. 1, pp. 109–118, 1990.
10. W. Ford, K. Xiang, W. Land, R. Congdon, Y. Li, and O. Sadik, "A Multi-class Probabilistic Neural Network for Pathogen Classification," *Procedia Computer Science*, vol. 20, pp. 348–353, 2013.
11. R. Wei, V. Gatterdam, R. Wieneke, R. Tampé and U. Rant, "Stochastic sensing of proteins with receptor-modified solid-state nanopores," Nature Nanotechnology, vol. 7, pp. 257-263,2012.
12. Walker, J. M. The protein protocols handbook; Springer Science &Business Media, 1996; Vol. 1996.
13. Domon, B.; Aebersold, R. Mass spectrometry and proteinanalysis. Science 2006, 312 (5771), 212−217.
14. Hellman, L. M.; Fried, M. G. Electrophoretic mobility shift assay (EMSA) for detecting protein−nucleic acid interactions. Nat. Protoc.2007, 2 (8), 1849−1861.
15. Dandliker, W.; De Saussure, V. Fluorescence polarization in immunochemistry. Immunochemistry 1970, 7 (9), 799−828.
16. (a) Engvall, E.; Perlmann, P. Enzyme-linked immunosorbentassay (ELISA) quantitative assay of immunoglobulin G. Immunochemistry1971, 8 (9), 871−874. (b)Van Weemen, B.; Schuurs, A.Immunoassay using antigen enzyme conjugates. FEBS Lett. 1971, 15(3), 232−236

17. Patolsky, F.; Zheng, G.; Hayden, O.; Lakadamyali, M.; Zhuang, X.; Lieber, C. M. Electrical detection of single viruses. Proceedings of the National Academy of Sciences 2004, 101 (39), 14017–14022.
18. Das, N.; Roychaudhuri, C. Reliability Study of Nanoporous Silicon Oxide Impedance Biosensor for Virus Detection: Influence of Surface Roughness. IEEE Transactions on Device and Materials Reliability 2015, 15 (3), 402–409.
19. 8. Stern, E.; Klemic, J. F.; Routenberg, D. A.; Wyrembak, P. N.; Turner-Evans, D. B.; Hamilton, A. D.; Lavan, D. A.; Fahmy, T. M.; Reed, M. A. Label-free immunodetection with CMOS-compatible semiconducting nanowires. Nature 2007, 445 (7127), 519–522.
20. Calvo, E. J.; Battaglini, F.; Danilowicz, C.; Wolosiuk, A.; Otero, M. Layer-by-layer electrostatic deposition of biomolecules on surfaces for molecular recognition, redox mediation and signal generation. Faraday Discussions 2000, 116, 47–65.