A Novel ML-based Symbol Detection Pipeline for Molecular Communication

Valerio Selis, Daniel Tunç McGuiness, and Alan Marshall, *Senior Member, IEEE*

***Abstract*—Molecular Communication (MC) is the process of sending information by the use of particles instead of electromagnetic (EM) waves. This change in paradigm allows the use of MC in areas where EM transmission is undesirable. These include underground, underwater and even intra-body communications. While this novel paradigm promises new areas for communication, one of the major setbacks is its relatively low throughput caused by the propagation speed. This can be improved by decreasing the symbol duration; however, this can be a detriment to the correct decoding of symbols. This paper proposes a novel symbol detection pipeline to increase the possible throughput without increasing the error rate of the communication. This is based on a machine-learning algorithm for classification tasks using an L-point discrete time moving average filter and a wide range of features. Extensive simulations with long sequences at different signal-to-noise ratio (SNR) values were performed to determine how well the proposed method detects symbols. The results show that our method can detect symbols received when On-Off Keying (OOK) modulations are used with a 10 dB gain, even when transmissions with untrained SNR values occur.**

***Index Terms*—Molecular communications, symbol detection, machine learning, signal processing.**

I. Introduction

MOLECULAR Communication (MC) is generally defined as the use of particles to convey information across a wide variety of distances [1]–[4]. This is a stark contrast to using electromagnetic (EM) or acoustic waves. This paradigm shifts from waves to particles allows novel implementations to areas where EM or acoustic would prove to be inefficient. These areas can include underwater [5]–[7], underground communication and infrastructure monitoring [8] where the wave (EM or acoustic) communication's poor performance was shown, such as strong attenuation when transmitting through the water-air boundary or its limited bandwidth [9]. For these environments, MCs have been studied as a possible alternative [10]–[12].

Another approach and a major attraction of MC is its possible use in biological systems [13]. Numerous studies were conducted for potential use in cell signalling applications, such as studying the process of calcium signalling [14]. It can also be used in nano-bot communication [15] to allow nano-scale robots to communicate with each other in in-vivo environments such as the human body [16], [17]. Of course, these are not the only areas in which MC is proposed to be useful, as it is also studied for possible use in other healthcare applications [18], [19], to study biological species such as microbial communities [20] and robotic communications [21].

An advantage of MC is, in itself, not bounded by any scale, whereas EM has physical limits within which a receiver/transmitter (i.e., an antenna [22]) can be constructed. This lower bound of EM has garnered interest in MC for use in micro- and nano-scale (nm - μm) environments with possible applications in intra-cellular communication [23], drug delivery systems [24], etc. These studies allowed further understanding of the various aspects of this novel communication paradigm such as testing established error correction methodologies such as Reed-Solomon codes [25], calculating the channel capacity of different mediums such as air or water, designing transmission protocols and initiating standardisation efforts [26] and, possible applications and examinations of security properties.

A significant hurdle that MC needs to overcome is its throughput speed or lack thereof. Unlike EM, where transmission is done comparable to the speed of light (c), MC relies on the use of particles. While it is currently a pipe dream to make MC comparable to EM in terms of speed, nevertheless, it can be considerably improved from its current standpoint. A major contributor to the poor throughput is the ISI (inter-symbol interference). This is caused by the physical constraints of the sensors used in detection or the environment itself where the propagation happens. As the transmission is done via particles, sensors physically interact with them to detect the received message. Due to this, the sensor needs to be given a time frame where it can remove the detected particle from the communication channel. If adequate time is not given, the channel will be saturated by particles from different transmissions. This would increase the background noise experienced by the sensor, which in turn would cause errors in decoding the received signal. However, if an appropriate method is used to decode the signal with a less than ideal time frame for removing the particles, the adverse effects of the increased background noise could be compensated, and a higher throughput may be achieved, which this paper proposes.

Authors in [27] have proposed a method to perform the symbol detection based on whether the last mass value received is below or above a static threshold (τ). We believe that a simple threshold detector is inefficient in decoding received symbols efficiently, especially in the presence of noisy channels.

The main contributions of the paper are as follows:

1) A novel pipeline to detect the symbols received using molecular communications is defined.

V. Selis and A. Marshall are with the Department of Electrical Engineering and Electronics, University of Liverpool, Brownlow Hill, L69 3GJ, Liverpool, UK, e-mail: V.Selis@liverpool.ac.uk, Alan.Marshall@liverpool.ac.uk. D. T. McGuiness is with the Department of Mechatronics, Management Centre Innsbruck (MCI), Austria, e-mail: Daniel.McGuiness@mci.edu.

2) An extensive number of long sequences of encoded data have been systematically used to train and test the proposed method.
3) A machine learning algorithm to improve the received symbol detection is developed and analysed.
4) Demonstrate the advantage of our solution compared to previous work.
5) The reliability and performance of the detection pipeline are validated when applied to unknown received sequences with different channel conditions.

The rest of this paper is organised as follows. Section II introduces the propagation theory behind the molecular communication used in this paper. In Section III, we describe and discuss a novel symbol detection pipeline. This is done by providing an analysis of the received signal and showing the steps required to train a machine learning algorithm for classification tasks. The results of the proposed symbol detection method are shown in Section IV. Finally, we present our conclusion and future work in Section V.

II. Molecular communication: A Theory of Propagation

A transmission involving particles, such as molecules, can be described using the generalised advection-diffusion equation (ADE) [28]. Depending on the environment studied, this equation is also known in the literature as convection–diffusion equation or generic (scalar) transport equation [29]. The 3-dimensional description is presented as follows:

$$\frac{\partial c}{\partial t} = \nabla \cdot (\mathbf{D}\nabla c) - \nabla \cdot (\mathbf{u}c) + K \,, \tag{1}$$

where c is the concentration of the mass in the environment (kg/m), t is the duration of the mass transfer process (s), $\mathbf{D}$ is the vector value for the coefficient of diffusivity (cm^2/s), $\mathbf{u}$ is the vector value of the velocity (i.e., advection process) (cm/s) and K is the sink and/or the source depending on the environment and its condition. If the system is a closed one, K is assumed negligible. For this work, it is assumed the environment possesses neither a sink nor a source ($K = 0$), there is no physical medium to guide the propagation (i.e., transmission is done in open air), and the change in the diffusion coefficient is deemed negligible ($\partial D/\partial t = 0$) during the propagation and absorption process and only the x-dimension is considered. Therefore, the boundary conditions for the 1-dimensional ADE are:

$$c(|x| > 0,\, t_0) = 0 \,, \tag{2a}$$
$$c(x = 0,\, t_0) = M_0\, \delta(x) \,, \tag{2b}$$
$$c(|x| \to \infty,\, t) = 0 \,, \tag{2c}$$

where M_0 is the initial mass injected into the environment (kg), x is the Cartesian propagation dimension, t_0 is the initial time (s) and $\delta(x)$ is the dimensional Dirac delta function. These conditions are known as the "*thin-film solution*" in the literature [30]. The solution with an unbounded domain can be expressed as:

$$c(x,\, t) = \frac{M_0}{\sqrt{(4\pi D_x t)}} \exp\left(-\frac{(x - u_x t)^2}{4 D_x t}\right) \,, \tag{3}$$

where u_x is the vectorial elements of $\mathbf{u}$ and D_x is the vectorial elements of $\mathbf{D}$ in the x-axis. To estimate the diffusion coefficient in a particular medium, the following equation is employed [31]:

$$D = \frac{2}{3}\sqrt{\frac{k_\mathrm{B}^3 T^3}{\pi^3}}\sqrt{\frac{1}{2m_\mathrm{A}} + \frac{1}{2m_\mathrm{B}}}\,\frac{4}{P(d_\mathrm{A} + d_\mathrm{B})^2} \,, \tag{4}$$

where k_B is the Boltzmann constant ($k_B = 1.380609 \times 10^{-23}\mathrm{J} \times \mathrm{K}^{-1}$), P is pressure, T is temperature, m_A, m_B and d_A, d_B are the molecular masses and diameters of chemical A and B respectively. One of these chemicals can be the signalling chemical and the other one is the chemical that is the medium (i.e., air), as the diffusion of the chemical depends on the medium it is propagating. Eq. (3) quantifies the concentration value of the sample in a given time (t) and space (x). The mass in a given transmission (θ) can be calculated by integrating the concentration function in the desired volume:

$$\theta = \int c\, dx \,. \tag{5}$$

The system has no sink/source ($K = 0$). Therefore, the particles used in the transmission process can either be in transmission (θ_T) or have been absorbed by the detector (θ_A). Both the aforementioned mass values *must* add up to the initial introduction of mass at the beginning of the transmission.

$$M_0 = \theta_T + \theta_A \,. \tag{6}$$

The absorbed mass (θ_A) (i.e., the transmission of bit 1) can be calculated simply by subtracting from the initial mass (M_0).

$$\theta_A = M_0 - \theta_T \,. \tag{7}$$

The particles present in the environment can be calculated by integrating the concentration function with respect to space. To calculate the chemicals absorbed by the detector (θ_A), the integration function is subtracted from the injected mass (M_0) [12], [32]–[35].

$$\theta_A(x,\, t) = M_0 - \int_{-x_\epsilon}^{+x_d} c(x,\, t)\, dx \,, \tag{8}$$

where x_d is the distance from the detector to the origin point ($x_0 = 0$) (m) and x_ϵ is the distance particles travel against the flow (m). This term is of diminutive value and can be treated as 0 ($x_\epsilon \approx 0$) if the system has an advection element. The solution to the integration given in Eq. (8) for transmission with no boundaries is given below:

$$\theta_1(x_d,\, x_\epsilon,\, t) = M_0 - \frac{M_0}{2}\left[\mathrm{erf}\left(\frac{x_d - u_x t}{\sqrt{4 D_x t}}\right) + \mathrm{erf}\left(\frac{x_\epsilon + u_x t}{\sqrt{4 D_x t}}\right)\right] \,, \tag{9}$$

where $\mathrm{erf}(x) = 2/\sqrt{\pi}\int_0^x e^{-t^2}\, dt$. The chemicals absorbed by the detector (θ_1) in a given period of T_s are given below:

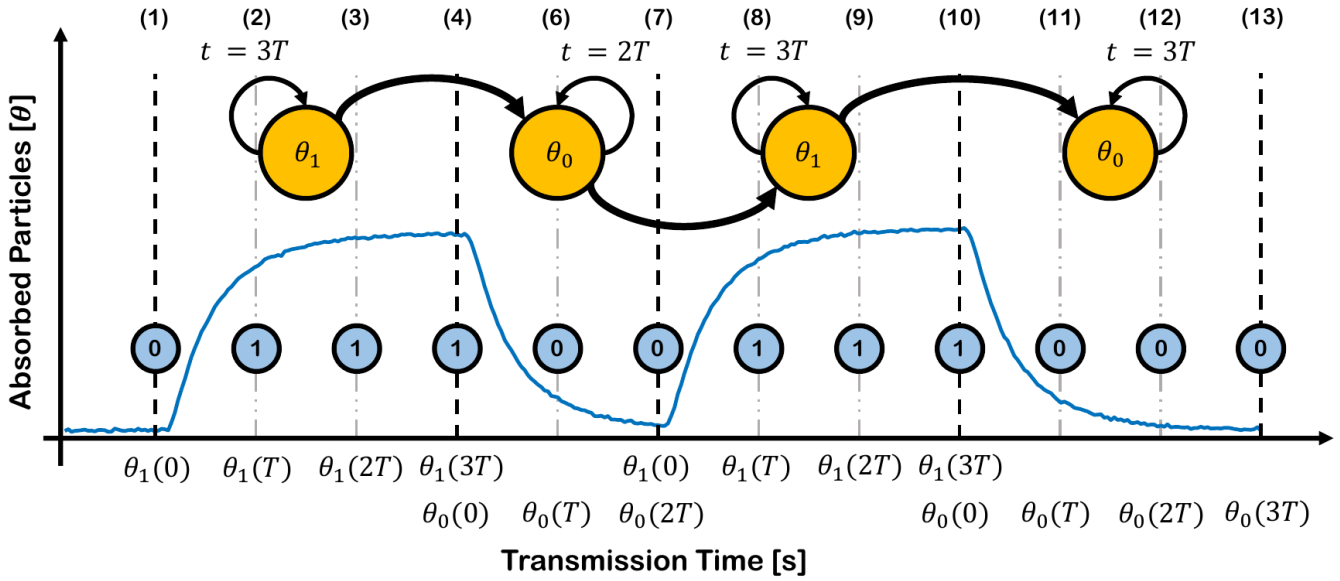

Fig. 1. A representative diagram of how the transmission is simulated with an example transmission of a bit sequence of 011100111000 with states of the transmission shown above the transmission. The first part of the simulation is to analyse the sequence based on the states. In this context, the states are defined as a bit value, which in this example are 0 and 1. In this example, there are five states which are 0-1-0-1-0 with durations of 1T -3T -2T -3T -3T. After this assessment, the system carries out the following procedures to initiate the simulation. In this example at time-point (1), the detector starts absorbing particles with the absorbing function θ_1 and this function continues until the time period of the state concludes at (4), in which the duration of the state is shown as the feedback loop to the state itself, with each bit-1 value having the absorbed mass value of θ_1(x, T), θ_1(x, 2T) and θ_1(x, 3T), respectively. When the time duration passes the time mark (4), the removal function (θ_0) initiates and starts removing the particles from the detector based on how many particles it has absorbed in the previous state [27].

$$M_R = \theta_1(x_d,\, 0,\, t = T_s) - \theta_1(x_d,\, 0,\, t = 0)\,. \tag{10}$$

Therefore, the removal of chemicals from the detector (θ_0) (i.e., the transmission of bit 0) to the outside environment can be expressed by the following expression.

$$\theta_0(x_d,\, x_\epsilon,\, t) = \frac{M_R}{2}\left[\mathrm{erf}\left(\frac{x_d - u_x t}{\sqrt{4D_x t}}\right) + \mathrm{erf}\left(\frac{x_\epsilon + u_x t}{\sqrt{4D_x t}}\right)\right]. \tag{11}$$

As shown in Eq. (9) and Eq. (11), the mass parameter is different in each equation: the former being the mass injected into the environment (M_0), and the latter being the mass absorbed by the detector (M_R). The theoretical model presented in this work, which was used to generate the transmission data, is based on the experimental work carried out in [32]–[35], where for the transmitter an odour generator is used and for the detector a membrane inlet mass spectrometer is used. This model assumes that there isn't any sensor saturation at the detector. A detailed diagram of its working is presented in Fig. 1, and additional details regarding the inner workings of the model are presented in [27].

A. Transmission Model and Coding Scheme

In this study, a channel model ($C\mathcal{H}$) in which open-air transmissions occur is used, as previously described. An On-Off Keying (OOK) modulation with an adapted Non-Return-to-Zero (NRZ) line code was implemented, where depending on the bit transmitted, either particles were introduced to the receiver (θ_1) or removed from the receiver (θ_0).

TABLE I
SIMULATION PARAMETERS

Simulation Parameter	Symbol	Value	Unit
Advective flow in x-axis	u_x	0.12	cm/s
Transmission Distance	x_d	1	cm
Diffusivity[1]	D	0.124	cm^2/s
Symbol Duration	T_s	20	s

[1] Modelled after diffusivity of acetone in laboratory conditions.

In the standard NRZ line code used in digital communications, there is a high voltage representing binary 1 and a low voltage representing binary 0 for the entire symbol period. In our adapted NRZ line code for MCs, there is an instantaneous injection of particles or an absence of particles' injection only at the beginning of the symbol period to represent binary 1 and 0, respectively. The parameters for the model can be seen in Table I. The noise is modelled as Additive White Gaussian Noise (AWGN) after the experimental validation of the noise characteristics presented in [32].

III. Symbol Detection Pipeline for Molecular Communications

The demodulation mechanism based on a static threshold for molecular-based transmissions may not perform well in specific environmental conditions, as previously identified. For this reason, a new pipeline based on a machine-learning algorithm is proposed in this paper. We have carried out several steps to obtain a trained machine learning algorithm that can be used as the basis of our symbol detection pipeline for MCs.

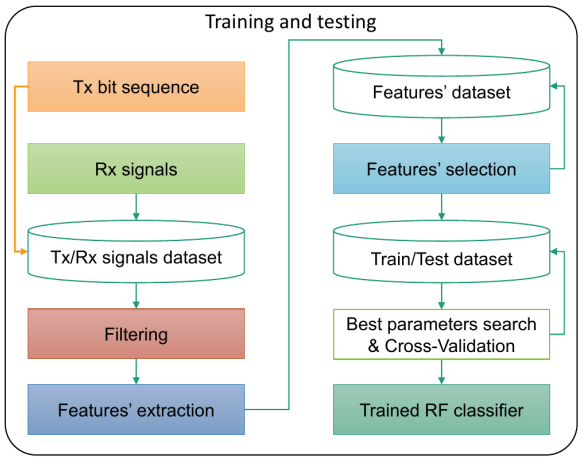

Fig. 2. Summary of the steps required to train and test the machine learning algorithm.

These steps consist of (i) creating the dataset containing the transmitted sequences and the corresponding received signal; (ii) the received signal is filtered; (iii) features are extracted from it; (iv) the extracted features are selected using their importance; and (v) the Random Forest classifier is trained and tested with different parameters. The steps are summarised in Fig. 2 and explained in more detail in the following subsections.

A. Tx sequences, Rx signals and Tx/Rx dataset

Random bit sequences have been generated to simulate several transmissions using MCs based on the OOK modulation with an adapted NRZ line code. The symbol duration or window size (T_s) is set to 20 seconds. The transmission process consists of generating binary symbols by instantaneously injecting (1-bit) or not injecting (0-bit) particles at the beginning of the symbol period. To detect the start of a transmission, a synchronisation bit (b^s) is used, which is identified by symbol 1. Let s_k be the k-th bit sequence, which is going to be transmitted using the OOK modulation, in which b_k^j is the j-th bit in s_k, then s_k can be represented as:

$$s_k = b_k^1, b_k^2, b_k^3, ..., b_k^n \,, \tag{12}$$

where n is the length of s_k and $b_k^1 = b^s$.

At the receiver, the signal is measured as the absorbed mass of the injected particles with a sampling rate of 1 second ($freq$ = 1 Hz). Therefore, the sampling of each transmitted symbol consists of $freq \cdot T_s = 20$ consecutive readings of the absorbed mass values from the sensor.

Let a_k be the k-th received signal, in which B_k^j is the j-th T_s absorbed mass values in a_k, then a_k can be represented as:

$$a_k = B_k^1, B_k^2, B_k^3, ..., B_k^n \,, \tag{13}$$

where n is the length of a_k. Let $M_{R_k}^{i,j}$ be the i-th absorbed mass value for the j-th T_s of a_k, for $i = 1, \cdots, T_s$ and $j = 1, \cdots, n$, then the transmitted j-th bit b_k^j of s_k (input I to the channel) can be represented as the received bit B_k^j of a_k (output O from the channel) as:

$$b_k^j \xrightarrow{I} \boxed{\mathcal{CH}} \xrightarrow{O} B_k^j = M_{R_k}^{1,j}, M_{R_k}^{2,j}, \cdots, M_{R_k}^{T_s \cdot j} \,. \tag{14}$$

Fig. 3 shows an example of a transmitted bit sequence containing the text "Hi" by using the parameters shown in Table I.

The text "Hi" is converted into binary by using the American Standard Code for Information Interchange (ASCII), and the b^s symbol is attached before its physical transmission, resulting in the following Tx bit sequence (S_{Hi}):

$$\begin{array}{rccccc} & b^s & b_{\text{Hi}}^2 & & \cdots & b_{\text{Hi}}^{17} \\ S_{\text{Hi}} = & 1 & 01001000 & & & 01101001 \end{array}$$

The bit sequence is then transmitted as shown in Fig. 3 (Tx signal) and simulated by using the method in [27]. As shown in this figure, the transmitted signal follows the OOK modulation with the adapted NRZ line code (left y-axis). In particular, for binary 1, a mass of 1 g is instantaneously injected into the environment, whereas for binary 0, there is the absence of particles' injection (0 g). At the receiver, a sensor is used to measure the absorbed mass (g) over time (right y-axis).

In Fig. 3, two examples of received signals are shown. One example of a received signal is when the signal-to-noise ratio (SNR) tends to infinity (Rx signal ideal), making this an ideal scenario as the transmitted mass is mostly absorbed over time. As expected, when there is a binary transmission of 1, the received signal increases as particles are injected into the environment and received over time. Whereas when there is a binary transmission of 0, the received signal decreases as the injected particles are absorbed by the receiver and/or evaporate/sublimate into the air. Moreover, when there is a transition stage from bit 1 to bit 0 or vice versa, the received signal fluctuates, making a peak or a valley, respectively. Around these fluctuations, the value of the absorbed mass changes rapidly. In the second example, the SNR value is set to 1 (Rx signal SNR 1), indicating that there is still more signal power than noise at the receiver. In this example, it can be seen that the received signal is difficult to be demodulated just by observing the absorbed mass values as these values highly fluctuate, also becoming negative due to the noise.

An initial dataset, called Tx/Rx dataset, has been created to analyse the received signals and obtain a trained machine learning algorithm. As per its name, this dataset contains the set S of transmitted bit sequences ($s_k \in S$) and the set A of the corresponding received signals ($a_k \in A$). Each s_k consists of n = 1 million consecutive bits, starting with b^s and followed by randomly generated 999,999 bits. The simulated received signals for different channel conditions based on SNR have been simulated using the method in [27] and the parameters in Table I. SNR ranged from -20 to 50 dB with a step value (T_{step}) of 5. For each SNR, the transmission of 10 bit sequences was simulated. Therefore, the Tx/Rx dataset contains a total of 150 transmitted bit sequences for a total of 150 million bits, and the correspondent 150 received signals formed by 3 billion absorbed mass values.

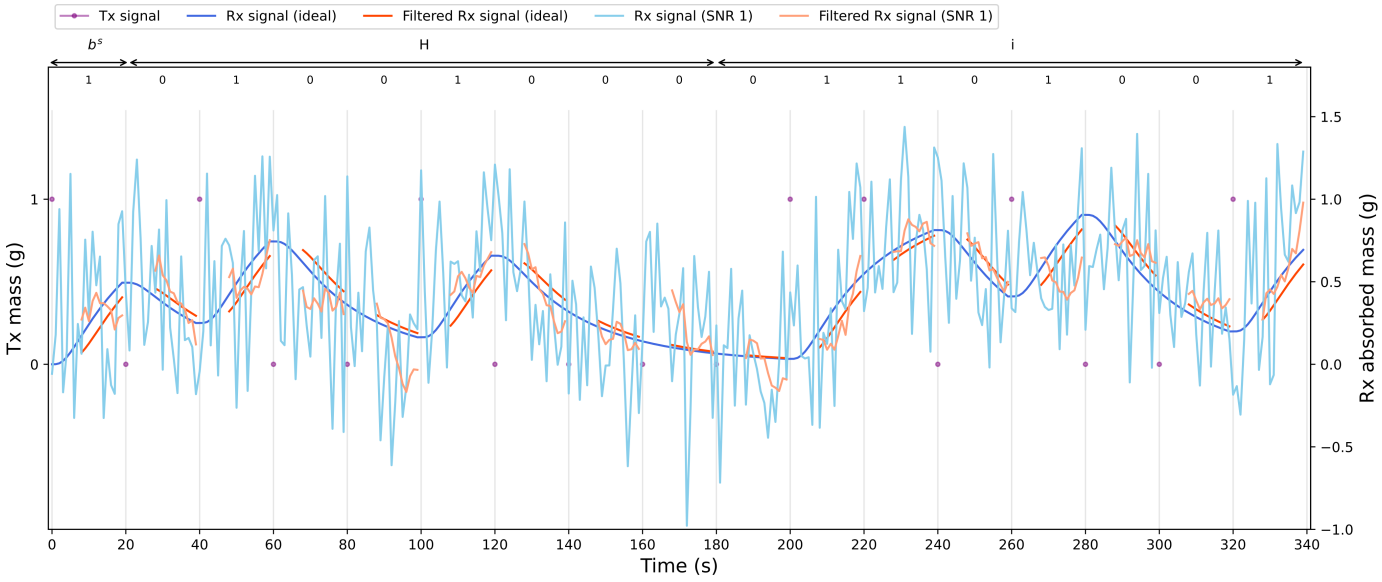

Fig. 3. Transmission (Tx) and reception (Rx) of the text "Hi" using molecular communications with the OOK modulation and the adapted NRZ line code. A binary 1 is represented by instantaneously injecting 1 g of mass, whereas a binary 0 is represented by the absence of particles' injection (0 g).

B. Filtering

When the SNR decreases, the received signal becomes noisy and difficult to be decoded, as can be seen in Fig. 3 (Rx signal SNR 1). For this reason, after receiving the signal, a filtering step has been introduced for each T_s to reduce the noise component. This step is achieved by applying an L-point discrete time moving average filter to the k-th received signal given by [36]:

$$fB_k^j[m] = \frac{1}{L}\sum_{i=0}^{L-1} B_k^j[m-i] \ , \tag{15}$$

where $B_k^j[]$ is the j received bit in the k-th received signal in input to the filter, L is the filter length and $fB_k^j[]$ is the average of L points in output from the filter. The number of points for each T_s decreases so that the filtered window size fT_s will be equal to $T_s - L$. This is a low-pass Finite Impulse Response (FIR) filter with an excellent time domain response that takes $L = T_s/4 = 5$ samples in input and produces as output a single value equal to the L-samples' average.

Let fa_k be the k-th filtered received signal, then it can be represented as:

$$fa_k = fB_k^1, fB_k^2, \cdots, fB_k^n \ , \tag{16}$$

where n is the length of s_k. Let $fM_{R_k}^{i\cdot j}$ be the i-th filtered absorbed mass value for the j-th fT_s of a_k, for $i = 1, \cdots, fT_s$ and $j = 1, \cdots, n$, then B_k^j can be represented as:

$$B_k^j \xrightarrow{I} \boxed{\mathcal{CH}} \xrightarrow{O} fB_k^j = fM_{R_k}^{1j}, fM_{R_k}^{2j}, \cdots, fM_{R_k}^{fT_s\cdot j} \ . \tag{17}$$

Moreover, filtering the signal will also cause a delay (d_f) in the filtered output signal of L seconds, as can be seen in Fig. 3 (Filtered Rx signal ideal and SNR 1). It is important to note that the filtering has been done per each T_s as this reflects what will happen in a real transmission in which a symbol is received every T_s seconds.

C. Features' extraction and Features' dataset

To characterise the filtered received signals and calculate the variability of each fB_k^j, n_{ms} statistical methods were used ($n_{ms} = 37$). These methods consisted of extracting features from the statistical and temporal domains from the filtered received signal in each fT_s. For each symbol, n_{ms} features were extracted, leading to a total of 5.85 billion values being generated for all the received signals present in the Tx/Rx dataset.

Let v_k be the feature vectors representing fa_k, V_k^j be the feature vector representing fB_k^j, and $e_k^{l\cdot j}$ be the l-th feature of the j-th V_k^j, for $l = 1, \cdots, n_{sm}$ and $j = 1, \cdots, n$, then:

$$fa_k \rightarrow v_k \ , \tag{18}$$

and a filtered received symbol can be represented as:

$$fB_k^j \rightarrow V_k^j = \left(e_k^{1j}, e_k^{2j}, ..., e_k^{n_{sm}\cdot j}\right) \ , \tag{19}$$

so that:

$$v_k = V_k^1, V_k^2, ..., V_k^n \ . \tag{20}$$

A new dataset is created containing $\{V, S\}$, called the Features' dataset, where V is the set containing all v_k. As there is a huge amount of data in this dataset, and not all the chosen features may be required to decode a symbol properly, a features selection step has been done. This allows to reduce the size of the Features' dataset and improve the machine learning algorithm performance. Moreover, by using fewer features, the demodulation phase of each symbol will be faster, and the computational resources required will be lower.

D. Features' selection and Train/Test dataset

In this work, Random Forest (RF) has been used for binary classification tasks, which is a collection of flow chart-like structures (forest) composed of nodes and leaves. Each node contains rules for the classification, which are obtained from the available features. At the same time, each leaf contains the resulting symbol, 0-bit or 1-bit. Several reasons have led to the use of RF; it can be used for binary and multiclass classification tasks, is computationally efficient, can handle high dimensional data, and can be used to select the most important features.

The Features' dataset has been used to perform 10 random permutations with a 70% training set and 30% testing set split. For each random permutation, the randomly chosen training set has been used to select the best p features. These have been obtained using the feature importance attribute from RF.

The ascending order rank of the feature importance has been calculated using the following:

$$rank\,(X) = \bigcup \begin{cases} rank\,(y_0 = n_{sm}) \\ rank\,(y_i) = rank\,(y_{i-1}) - 1 \end{cases}; \quad \text{for } i = 1, ..., n\,, \quad (21)$$

where X is an ascending sorted set with the sum of how many times each feature is selected, y_i is the i-th element in X and n is the number of elements in X. The rank value of n_{sm} will represent the best feature chosen, and the rank value of 1 will represent the worst feature. Fig. 4 shows for each feature the value obtained from the ascending order of its rank for each SNR value. Moreover, features have been grouped into three main ranges: the first range highlighted in yellow-orange shades with SNR values less than 0 dB [-20 dB, 0 dB); the second range highlighted in green shades with SNR values between 0 dB and 15 dB [0 dB, 15 dB]; third range highlighted in blue shades with SNR values above 15 dB (15 dB to 50 dB]. These ranges have been helpful in understanding which features can be used when there is more noise than signal power (first SNR range), when there is slightly more signal power than noise (second SNR range) and when there is high signal power than noise (third SNR range).

From the figure, it is possible to see that the best feature is the upper bound (U_b), whereas the worst feature is the count of values higher than the mean (count > μ). Moreover, the mean (μ) is the best feature for SNR less than 0 dB, whereas the root mean square (RMS) is the best feature for SNR greater than and equal to 0 dB. Despite this, RMS is not a good feature to be used within the first SNR range, making its use for classification not as good as other features which have obtained better overall importance across all SNR ranges, such as $slope$, last filtered window value ($fM_{R_k}^{fT_s \cdot j}$), $median$ and mean change (μ_{change}). The last filtered window value is a better feature than the upper bound feature for the second range. Furthermore, the $slope$ feature is slightly better than the upper bound feature for the third range. Table II shows the best $p = 14$ features used with RF.

TABLE II
BEST $p = 14$ FEATURES [37], [38].

Feature	Equation
Mean	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
Last filtered window value	$fM_{R_k}^{fT_s \cdot j}$
Slope	$slope = \text{polyfit}(y, X, 1)\,[0]$
Root mean square	$RMS = \sqrt{\frac{1}{N}\sum_i x_i^2}$
Upper bound	$U_b = \mu + \sigma$
Median	$\tilde{x} = \frac{1}{2}\left(x_{\frac{N}{2}} + x_{\frac{N}{2}+1}\right)$ of $sort(X)$
Mean change	$\mu_{change} = \frac{1}{1-N}\,(x_N - x_1)$
Mean diff	$\mu_{diff} = \frac{1}{N}\sum_{i=1}^{N-1}(x_{i+1} - x_i)$
Absolute energy	$E = \sum_{i,...,n} x_i^2$
Autocorrelation	$c_{xx}(k) = \sum_{n=0}^{N-k-1} x_{n+k} \cdot x_n^*$
Maximum value	$\max(X)$
Lower bound	$L_b = \mu - \sigma$
Peak to peak distance	p–p = $\lvert\max(X) - \min(X)\rvert$
Absolute maximum	$f(x_M) = \lvert\max(X)\rvert$

X = all samples; x_i = i-th sample;
N = number of samples; $sort$ = values in sorted order;
$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$; $y = [0, \cdots, fr_s]$.

Let sv_k be all feature vectors with p selected features representing fa_k, SV_k^j be the selected feature vector representing fB_k^j, and $se_k^{p \cdot j}$ be the p-th feature of the j-th SV_k^j, then:

$$fa_k \rightarrow sv_k\,, \quad (22)$$

and a filtered received symbol can be expressed as:

$$fB_k^j \rightarrow SV_k^j = \left(se_k^{3j}, se_k^{4j}, ..., se_k^{p \cdot j}\right)\,, \text{ for } p = 3, \cdots, 14\,. \quad (23)$$

Using the best p features, a dimensionality reduction of the Features' dataset is achieved, and this will also reduce the time required to detect the symbol. New datasets were created containing $\{SV, S\}$, called the Train/Test datasets, where SV is the set containing all sv_k which are the samples, and S contains all s_k which are the classes. The Train/Test datasets were created by selecting the p best features, as shown in Table III.

E. Best parameters search & Cross-Validation

The last important step, which is used as the core of our proposed solution, consists of training RF classifiers that can be used to detect the symbols. As handling the Train/Test datasets still require a lot of memory, a randomised grid search has been used to select the hyperparameters for each classifier.

This step has been done by splitting the Train/Test datasets into a 70% training set and 30% testing set split. A stratified KFold cross-validation method has been used to avoid underfitting and overfitting the classifiers to the training set. It randomly splits the training set into K parts, and each part

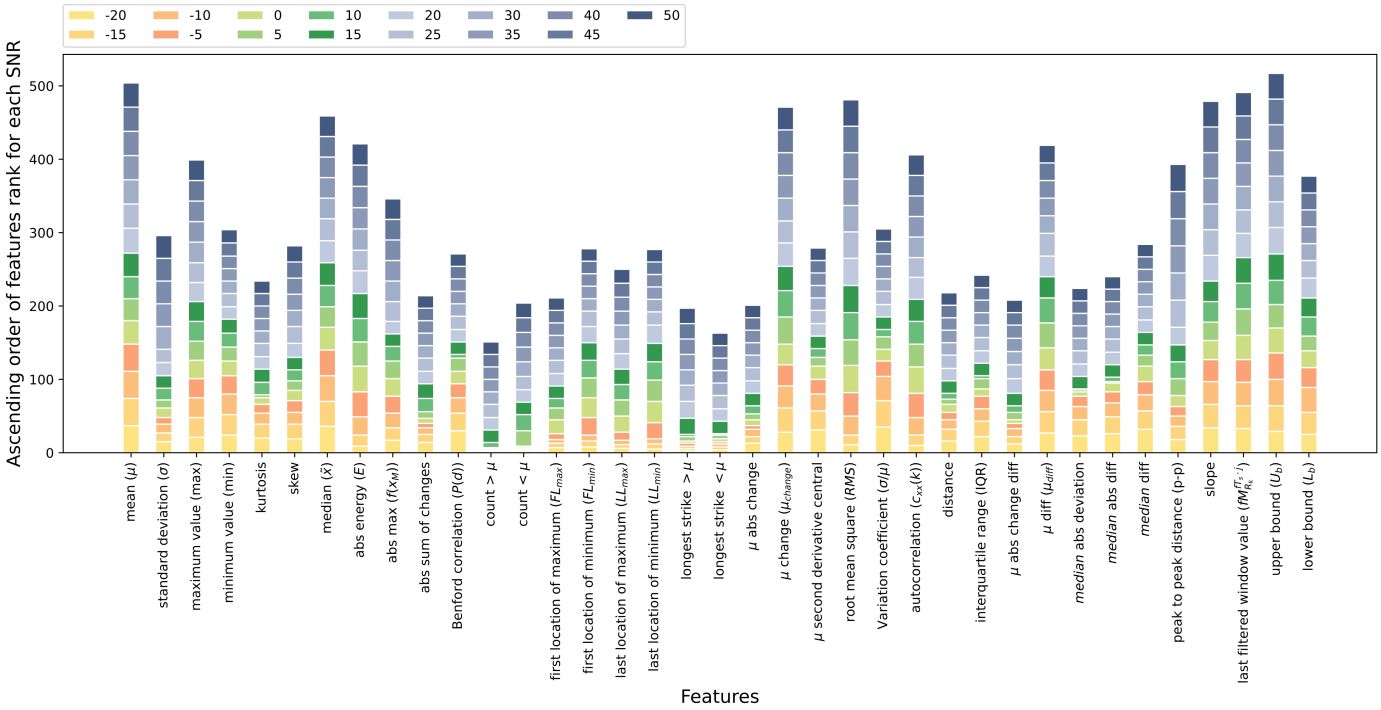

Fig. 4. Ascending order rank of the feature importance values for each SNR.

TABLE III
SELECTED FEATURES TO CREATE TRAIN/TEST DATASETS.

Feature	Equation
$p = 1$	μ
$p = 2$	$p = 1$ feature and $f_{m_k^{fT_s \cdot j}}$
$p = 3$	$p = 2$ features and $slope$
$p = 4$	$p = 3$ features and RMS
$p = 5$	$p = 4$ features and U_b
$p = 6$	$p = 5$ features and $\tilde{x}$
$p = 7$	$p = 6$ features and μ_{change}
$p = 8$	$p = 7$ features and μ_{diff}
$p = 14$	$p = 8$ features, E, $c_{xx}(k)$, max, L_b, p–p and $f(x_M)$

TABLE IV
RANDOM FOREST CLASSIFIER HYPERPARAMETERS [39].

Parameter	Value
n_estimators	10, 20, **50***, 100, 200, 250, 500, 1000, 1500, 2000, 5000
criterion	gini, **entropy***
max_features	auto, **sqrt***, log2
max_depth	None, **10***, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110
min_samples_split	2, **5***, 10
min_samples_leaf	1, **2***, 4
bootstrap	**False***, True
*** best hyperparameters.**	

has the same quantity of the best p features per symbol. Successively, $K-1$ parts are used by the classifier to learn, and one part is used to test how well it can decode each symbol. This process is repeated K times to retrieve the best parameters, which gave the highest classification accuracy; in this case, $K = 3$. The hyperparameters used and the best selected hyperparameters (highlighted in bold) are shown in Table IV.

The trained RF classifiers with the chosen hyperparameters are then used to assign each fB_k^j described by the best p features to a specific symbol, 1-bit or 0-bit.

F. Application to Molecular Communications

In a real scenario in which our OOK modulation will be used for MCs, the transmitter needs to inject particles at the beginning of the transmission, representing the synchronisation bit. Then every 20 seconds, the transmitter needs to inject particles only if the symbol 1-bit needs to be transmitted. The receiver demodulates the signal received after sampling 20 measured values gathered from the sensor. The received signal is obtained by measuring the absorbed mass values for each window size, as explained in Subsection III-A (Stage 1). Each sampled received signal for the given window size is then filtered using the 5-point discrete time moving average filter described in Subsection III-B (Stage 2). Following the filtering, the best p features obtained in Subsection III-D are extracted (Stage 3). Finally, the trained RF classifiers obtained in Subsection III-E are used to estimate the transmitted bit ($\mathbb{E}[b]$) by using the extracted features (Stage 4). Algorithm 1 shows the pseudo-code for the proposed symbol detection pipeline with the highlighted four stages, where:

- read_sensor() is a blocking function that reads an ab-

sorbed mass value from the sensor every second.
- lfilter() is the function that implements the 5-point discrete time moving average filter.
- best_p_features() is the function that extracts the best p features.
- trained_RF() is the function that implements the trained RF classifier with the chosen hyperparameters and p features.

Algorithm 1 Pseudo-code for the symbol detection algorithm used for molecular communications.

Input: n (length of the sequence), T_S (bit duration)
Output: $\mathbb{E}[b^j]$ (estimated transmitted bit)

for $j = 1$ to n **do**
  $B^j \leftarrow$ new List // Stage 1
  for $i = 1$ to T_S **do**
    $M_R^{i \cdot j} \leftarrow$ read_sensor()
    Append $M_R^{i \cdot j}$ to B^j
    $i \leftarrow i + 1$
  end for
  $fB^j \leftarrow$ lfilter(B^j) // Stage 2
  $SV^j \leftarrow$ best_p_features(fB^j) // Stage 3
  $\mathbb{E}[b^j] \leftarrow$ trained_RF(SV^j) // Stage 4
end for

IV. Simulation and Results

To test the proposed symbol detection pipeline, a new dataset has been created, called the Unknown dataset. As per its name, this dataset contains the unknown set US of simulated bit sequences ($us_k \in US$) and the set UA of the corresponding unknown received signals ($ua_k \in UA$). As for the Tx/Rx dataset, each us_k consists of $n = 1$ million consecutive bits, starting with b^s and followed by randomly chosen 999,999 bits. The simulated received signals for SNR ranging from -20 to 50 dB with T_{step} equal to 2.5 have been generated using the method in [27]. It is important to note that the T_{step} value used for generating the Tx/Rx dataset was equal to 5. Therefore, in the Unknown dataset, there are received signals which may be completely different from the signals used to train the RF classifiers, e.g. received signals simulated with SNR = -17.5 dB were not included in the Tx/Rx dataset. The Unknown dataset contains 290 transmitted bit sequences (us_k) and the corresponding 290 received signals (ua_k). Despite having the transmit bit sequences inside the dataset, these are used only to check whether the estimated received bits are correct or not. Algorithm 1 is then used to estimate the unknown transmitted bits ($\mathbb{E}[ub_k^j]$) in UA for different p features. Simulations have been performed by implementing Algorithm 1 in Python 3 using the NumPy [37], SciPy [40], Scikit-learn [39] and tsfresh [38] modules.

A us_k for a specific SNR value is used to simulate a real transmission in MCs. Each value in the correspondent ua_k is then given in input to Algorithm 1 to simulate the read_sensor() function. The proposed pipeline returns each $\mathbb{E}[ub_k^j]$ for the given ua_k. The estimated received bits are then compared with the transmitted bits in us_k to determine how well the proposed method detects received symbols.

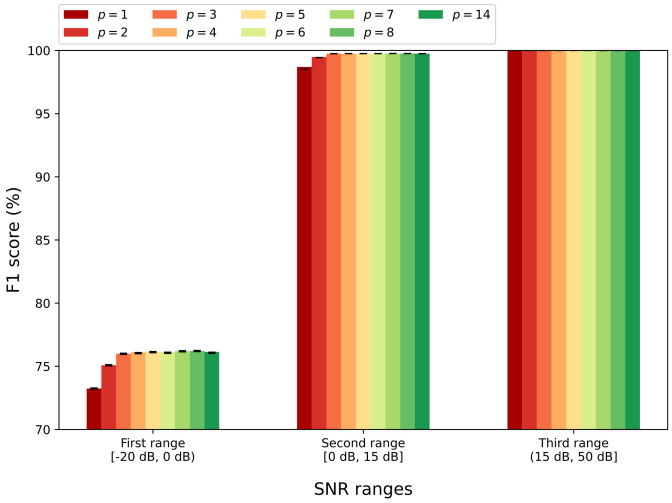

Fig. 5. F1 score of RF classifiers for each p best features and SNR ranges.

This is a binary classification problem with two classes: 0-bit symbol (positive class), and 1-bit symbol (negative class). When the estimated values $\mathbb{E}[ub_k^j]$ are compared with the real values ub_k^j, a confusion matrix can be created where:

- True Positives (TP): a 0-bit is classified as a 0-bit (correct result);
- True Negatives (TN): a 1-bit is classified as a 1-bit (correct absence of result);
- False Positives (FP): a 1-bit is classified as a 0-bit (unexpected result);
- False Negatives (FN): a 0-bit is classified as a 1-bit (missing result).

To evaluate how the classifiers performed, the balanced F-score (F1 score) performance measure has been used, which is the harmonic mean of the precision and recall metrics, where:

- Precision measures how many samples are classified as positive and are actually positive $\left(P = \frac{TP}{TP+FP}\right)$;
- Recall measures how many positive samples are classified as positive $\left(R = \frac{TP}{TP+FN}\right)$;

so that the F1 score can be calculated as:

$$F1 = 2\frac{P \times R}{P + R} \,. \tag{24}$$

The F1 score performance measure gives equal importance to FP and FN, and it is useful when the datasets are unbalanced, which may be the case as bits were randomly generated and the b_s is always 1. Results from this performance measure are between 0 (worst result) and 1 (best result). Fig. 5 shows the performance of RF classifiers for each p best features and SNR ranges. As it can be seen in this figure, there is a plateau when the number of best p features selected is equal to three. When $p \geq 3$, the F1 score value is around 76%, 99.75% and 100% for the first, second and third SNR groups, respectively. The results obtained show that the proposed RF classifiers will classify bits received with an SNR value greater and equal to 0, whereas it may struggle to classify bits received with an SNR value less than 0.

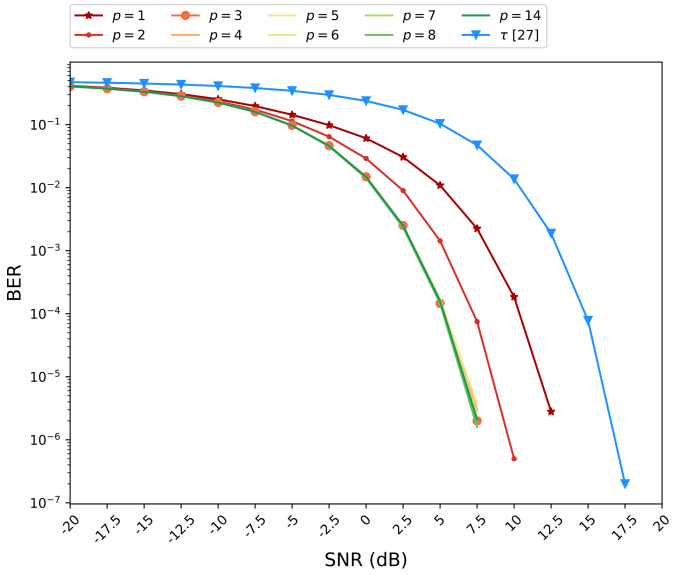

Fig. 6. Simulation results showing the bit error rate for each signal-to-noise ratio value of the proposed RF classifiers with the best p features and the method proposed in [27].

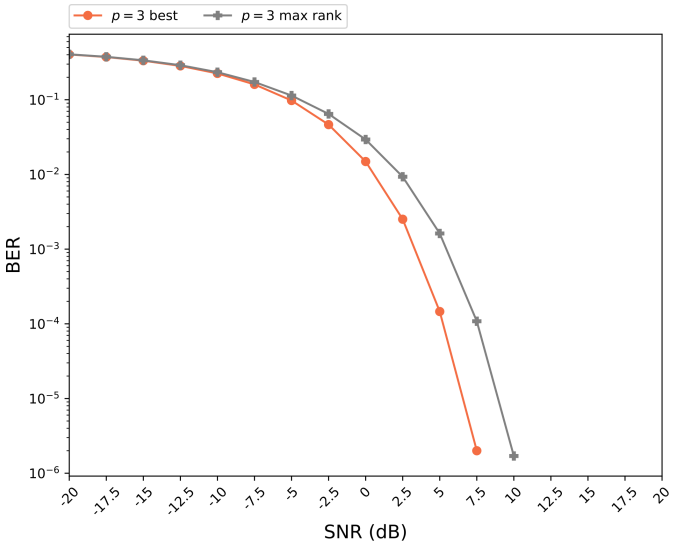

Fig. 7. Simulation results showing the bit error rate for each signal-to-noise ratio value of the proposed RF classifier with the best $p = 3$ features ($p = 3$ best) and the RF classifier with the $p = 3$ features having the maximum rank values ($p = 3$ max rank).

As we are interested in determining how well the proposed pipeline detects the received symbols, the bit error rate (BER) measure is used. This is commonly used for digital transmissions, and it is calculated by using the following equation:

$$BER = \frac{N_{errors}}{N_{bits}} = \frac{FP + FN}{TP + TN + FP + FN}, \tag{25}$$

where N_{errors} is the number of bits that were wrongly received and N_{bits} is the number of transmitted bits.

Results of the proposed RF classifiers compared with the method proposed in [27] are shown in Fig. 6. This shows the high efficiency of our method in terms of bit detection performance compared to the method proposed in [27]. As expected, when the best $p \leq 2$ features are used, the performance of the proposed RF classifiers is not the best that can be reached. The best performance can be achieved using the RF classifier trained with only the best $p = 3$ features. This was also expected as we had a plateau in Fig. 5 for that number of selected features. When the best $p = 3$ features are used, the best performance is also achieved for SNR values less than 0, where the F1 score for the first SNR range was lower than the other two SNR ranges.

As discussed in Subsection III-D, SNR ranges are helpful in understanding which features can be used. To further evaluate their importance in selecting the best p features, a comparison between the best $p = 3$ features ($p = 3$ best) and the $p = 3$ features having the maximum rank values ($p = 3$ max rank) has been performed. Results are shown in Fig. 7 and it is possible to observe that the proposed RF classifier using the best $p = 3$ features (μ, $f_{m_k^{fT_s} \cdot j}$ and $slope$) outperform the RF classifier using the $p = 3$ features having the maximum rank values (U_b, μ and $f_{m_k^{fT_s} \cdot j}$).

V. Conclusion

In this paper, we have presented a novel pipeline to detect symbols received during MCs. This has been evaluated by simulating the transmission of long sequences composed of 1 million bits using the OOK modulation. Simulated transmissions for SNR values between -20 and 50 dB with a 5 dB step were used for creating a reference dataset for training RF classifiers. Successively, a new dataset composed of simulated transmissions for the same SNR range with a 2.5 dB step was used for creating an unknown dataset for testing. The final pipeline uses three features from the filtered received mass values in each window: the mean, the last received value and the slope. This allows the detection of bits received with a 10 dB gain compared to the previous solution, even when completely unknown transmissions occur. This shows that the proposed pipeline is reliable even when there are variable SNR values during a transmission. Future research will involve the adoption of this pipeline with different model parameters, more than one compound used to transmit bits and the use of different line codes like Manchester to transmit other types of information.

References

[1] N. Farsad, H. B. Yilmaz, A. Eckford, C.-B. Chae, and W. Guo, "A comprehensive survey of recent advancements in molecular communication," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1887–1919, 2016.

[2] B. Atakan, *Molecular Communications and Nanonetworks*. Springer, 2016.

[3] N. Farsad and A. Goldsmith, "Neural network detection of data sequences in communication systems," *IEEE Transactions on Signal Processing*, vol. 66, no. 21, pp. 5663–5678, 2018.

[4] N. Farsad, D. Pan, and A. Goldsmith, "A novel experimental platform for in-vessel multi-chemical molecular communications," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 2017, pp. 1–6.

[5] J. Partan, J. Kurose, and B. N. Levine, "A survey of practical issues in underwater networks," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 11, no. 4, pp. 23–33, 2007.

[6] H. Riksfjord, O. T. Haug, and J. M. Hovem, "Underwater acoustic networks-survey on communication challenges with transmission simulations," in *2009 Third International Conference on Sensor Technologies and Applications*. IEEE, 2009, pp. 300–305.

[7] W. W. Au, P. E. Nachtigall, and J. L. Pawloski, "Acoustic effects of the atoc signal (75 hz, 195 db) on dolphins and whales," *The Journal of the Acoustical Society of America*, vol. 101, no. 5, pp. 2973–2977, 1997.

[8] F. Stajano, N. Hoult, I. Wassell, P. Bennett, C. Middleton, and K. Soga, "Smart bridges, smart tunnels: Transforming wireless sensor networks from research prototypes into robust engineering infrastructure," *Ad Hoc Networks*, vol. 8, no. 8, pp. 872–888, 2010.

[9] X. Che, I. Wells, G. Dickers, P. Kear, and X. Gong, "Re-evaluation of rf electromagnetic communication in underwater sensor networks," *IEEE Communications Magazine*, vol. 48, no. 12, pp. 143–151, 2010.

[10] W. Guo, I. Atthanayake, and P. Thomas, "Vertical underwater molecular communications via buoyancy: Gaussian velocity distribution of signal," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–6.

[11] O. Yetimoğlu, A. Dilmaç, Z. C. Canbek, and H. B. Yilmaz, "Underwater testbed for molecular communication," in *2021 29th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2021, pp. 1–4.

[12] S. Qiu, W. Guo, S. Wang, N. Farsad, and A. Eckford, "A molecular communication link for monitoring in confined environments," in *2014 IEEE International Conference on Communications Workshops (ICC)*. IEEE, 2014, pp. 718–723.

[13] T. Suda and T. Nakano, "Molecular communication as a biological system," in *2018 IEEE International Conference on Sensing, Communication and Networking (SECON Workshops)*. IEEE, 2018, pp. 1–4.

[14] P. He, T. Nakano, D. Wu, B. Yang, H. Liu, and X. Han, "Calcium signaling in mobile molecular communication networks," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[15] N. Farsad, "Molecular communication: Interconnecting tiny nanobio devices," *GetMobile: Mobile Computing and Communications*, vol. 22, no. 2, pp. 5–10, 2018.

[16] D. T. McGuiness, V. Selis, and A. Marshall, "Molecular-based nano-communication network: A ring topology nano-bots for in-vivo drug delivery systems," *IEEE Access*, vol. 7, pp. 12 901–12 913, 2019.

[17] L. Lin, W. Li, R. Zheng, F. Liu, and H. Yan, "Diffusion-based reference broadcast synchronization for molecular communication in nanonetworks," *IEEE Access*, vol. 7, pp. 95 527–95 535, 2019.

[18] Y. Cevallos, L. Tello-Oquendo, D. Inca, D. Ghose, A. Z. Shirazi, and G. A. Gomez, "Health applications based on molecular communications: A brief review," in *2019 IEEE International Conference on E-health Networking, Application & Services (HealthCom)*. IEEE, 2019, pp. 1–6.

[19] I. F. Akyildiz, M. Pierobon, and S. Balasubramaniam, "Moving forward with molecular communication: From theory to human health applications [point of view]," *Proceedings of the IEEE*, vol. 107, no. 5, pp. 858–865, 2019.

[20] L. Guo, X. He, and W. Shi, "Intercellular communications in multispecies oral microbial communities," *Frontiers in microbiology*, vol. 5, p. 328, 2014.

[21] B. Atakan, "Molecular communication among nanomachines," in *Molecular Communications and Nanonetworks*. Springer, 2014, pp. 1–24.

[22] W. Davis, T. Yang, E. Caswell, and W. Stutzman, "Fundamental limits on antenna size: a new limit," *IET microwaves, antennas & propagation*, vol. 5, no. 11, pp. 1297–1302, 2011.

[23] Y. Moritani, S. Hiyama, S. Nomura, K. Akiyoshi, and T. Suda, "A communication interface using vesicles embedded with channel forming proteins in molecular communication," in *2007 2nd Bio-Inspired Models of Network, Information and Computing Systems*. IEEE, 2007, pp. 147–149.

[24] U. A. Chude-Okonkwo, R. Malekian, B. T. Maharaj, and A. V. Vasilakos, "Molecular communication and nanonetwork for targeted drug delivery: A survey," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 3046–3096, 2017.

[25] M. B. Dissanayake, Y. Deng, A. Nallanathan, M. Elkashlan, and U. Mitra, "Interference mitigation in large-scale multiuser molecular communication," *IEEE Transactions on Communications*, vol. 67, no. 6, pp. 4088–4103, 2019.

[26] Y. Sangar and B. Krishnaswamy, "Link layer protocol for molecular communication networks," in *Proceedings of the Sixth Annual ACM International Conference on Nanoscale Computing and Communication*, 2019, pp. 1–6.

[27] D. T. McGuiness, S. Giannoukos, A. Marshall, and S. Taylor, "Modulation analysis in macro-molecular communications," *IEEE Access*, vol. 7, pp. 11 049–11 065, 2019.

[28] T. Stocker, *Introduction to climate modelling*. Springer Science & Business Media, 2011.

[29] C. E. Baukal Jr, V. Gershtein, and X. J. Li, *Computational fluid dynamics in industrial combustion*. CRC press, 2000.

[30] J. Crank, *The mathematics of diffusion*. Oxford university press, 1979.

[31] E. H. Kennard *et al.*, *Kinetic theory of gases*. McGraw-hill New York, 1938, vol. 483.

[32] D. T. McGuiness, S. Giannoukos, A. Marshall, and S. Taylor, "Experimental results on the open-air transmission of macro-molecular communication using membrane inlet mass spectrometry," *IEEE Communications Letters*, vol. 22, no. 12, pp. 2567–2570, 2018.

[33] ——, "Parameter analysis in macro-scale molecular communications using advection-diffusion," *IEEE Access*, vol. 6, pp. 46 706–46 717, 2018.

[34] D. T. McGuiness, S. Giannoukos, S. Taylor, and A. Marshall, "Experimental and analytical analysis of macro-scale molecular communications within closed boundaries," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 5, no. 1, pp. 44–55, 2019.

[35] ——, "Analysis of multi-chemical transmission in the macro-scale," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 6, no. 2, pp. 93–106, 2020.

[36] S. W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing*. USA: California Technical Publishing, 1997.

[37] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2

[38] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package)," *Neurocomputing*, vol. 307, pp. 72–77, 2018.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[40] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

Valerio Selis was born in Cagliari, Sardinia, Italy, in 1983. He received the M.Sc. degree in Computer Science from the University of Cagliari, Italy, and the Ph.D. in Electrical Engineering and Electronics from the University of Liverpool (UoL), U.K. Currently, he is working as a Lecturer with the Advanced Networks Research Group at UoL. His recent research was focused on molecular-based nano-communication networks for in-vivo drug delivery systems, specifically nano-machine to nano-machine communications. Moreover, he has been Product Development Director at Traffic Observation via Management Ltd. His research interests include machine learning, molecular communications, nano-networks, trust management and Internet of Things.

Daniel Tunç McGuiness received the B.Sc. degree in electrical engineering from Istanbul Technical University (ITU), Turkey, and the Ph.D. degree from the University of Liverpool (UoL), U.K., where he studied macro-scale molecular communications both experimentally and theoretically. He took a final-year specialisation in solid-rotor induction motors with ITU. He is currently a Lecturer at Management Centre Innsbruck (MCI). His focuses are on molecular and nano-communication systems, and electric machines.

Alan Marshall holds the chair in Communications Networks at the University of Liverpool where he is director of the Advanced Networks Group and Head of Department. He is a Fellow of The Institution of Engineering and Technology. He has spent over 30 years working in the Telecommunications and Defense Industries. He has been visiting professor in network security at the University of Nice/CNRS, France, and Adjunct Professor for Research at Sunway University Malaysia. He has published over 200 scientific papers and holds a number of joint patents in the areas of communications and network security. He has formed a successful spin-out company Traffic Observation and Management (TOM) Ltd. His research interests include Network architectures and protocols; Mobile and Wireless networks; Network Security; high-speed packet switching, QoS/QoE architectures; and Multi-Sensory Communications including haptics and olfaction.