



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0036948
(43) 공개일자 2023년03월15일

(51) 국제특허분류(Int. Cl.)

G06F 21/56 (2013.01)

(52) CPC특허분류

G06F 21/562 (2013.01)

G06F 21/561 (2013.01)

(21) 출원번호 10-2021-0163930

(22) 출원일자 2021년11월25일

심사청구일자 2021년11월25일

(30) 우선권주장

1020210119714 2021년09월08일 대한민국(KR)

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

박상민

경기도 남양주시 와부읍 덕소로 270, 106동 1001호(한강우성아파트)

김휘강

서울특별시 서초구 서운로 104, 201동 804호(서초동, 래미안서초에스티지에스)

전상훈

경기도 수원시 장안구 이목로 24, 106동 2102호(정자동, 수원 SK SKY VIEW)

(74) 대리인

이대호, 박건홍

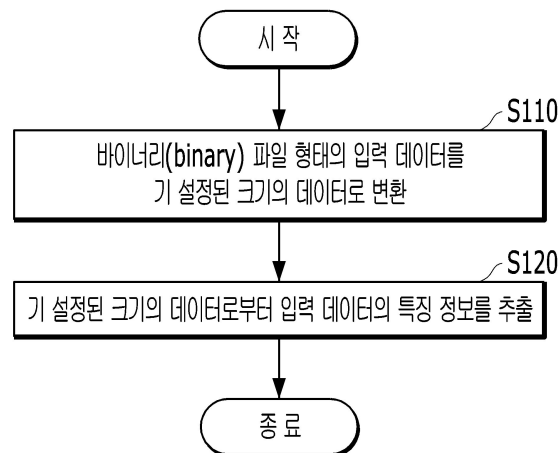
전체 청구항 수 : 총 16 항

(54) 발명의 명칭 바이너리 파일 고정화와 웨이블릿 신호처리를 통한 바이너리 파일 특징 정보 추출

(57) 요약

본 개시의 몇몇 실시예에 따른, 적어도 하나의 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 파일 특징 정보를 추출하는 방법이 개시된다. 상기 파일 특징 정보를 추출하는 방법은 바이너리(binary) 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환하는 단계; 및 상기 기 설정된 크기의 데이터로부터 상기 입력 데이터의 특징 정보를 추출하는 단계를 포함할 수 있다.

대표도 - 도2



명세서

청구범위

청구항 1

적어도 하나의 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 파일 특징 정보를 추출하는 방법으로서,
바이너리(binary) 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환하는 단계; 및
상기 기 설정된 크기의 데이터로부터 상기 입력 데이터의 특징 정보를 추출하는 단계;
를 포함하는,
파일 특징 정보를 추출하는 방법.

청구항 2

제 1 항에 있어서,
상기 바이너리(binary) 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환하는 단계는,
히든 마르코프 모델(hidden Markov model)을 이용하여, 상기 입력 데이터를 상기 기 설정된 크기의 데이터로 변환하는 단계;
를 포함하는,
파일 특징 정보를 추출하는 방법.

청구항 3

제 1 항에 있어서,
상기 바이너리(binary) 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환하는 단계는,
상기 입력 데이터를 기억장치의 크기를 나타내는 제 1 단위(unit)로 구분하여 인식하는 단계;
상기 제 1 단위로 인식된 데이터의 개수를 카운팅하는 단계; 및
상기 제 1 단위의 데이터의 개수에 기초하여, 상기 입력 데이터를 상기 기 설정된 크기의 데이터로 변환하는 단계;
를 포함하는,
파일 특징 정보를 추출하는 방법.

청구항 4

제 1 항에 있어서,
상기 바이너리(binary) 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환하는 단계는,
상기 입력 데이터를 바이트 단위로 인식하는 단계;
인식된 복수의 바이트들 각각의 개수를 카운팅하는 단계; 및
상기 복수의 바이트들 각각의 개수에 기초하여, 상기 입력 데이터를 상기 기 설정된 크기의 데이터로 변환하는 단계;

를 포함하는,
파일 특징 정보를 추출하는 방법.

청구항 5

제 4 항에 있어서,
상기 복수의 바이트들 각각의 개수에 기초하여, 상기 입력 데이터를 상기 기 설정된 크기의 데이터로 변환하는 단계는,
상기 복수의 바이트들 각각의 개수에 기초하여, 전체 데이터 중 상기 복수의 바이트들 각각의 출현 확률을 계산하는 단계; 및
상기 출현 확률에 기초하여, 상기 입력 데이터를 상기 기 설정된 크기의 데이터로 변환하는 단계;
를 포함하는,
파일 특징 정보를 추출하는 방법.

청구항 6

제 1 항에 있어서,
상기 바이너리(binary) 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환하는 단계는,
상기 입력 데이터를 바이트 단위로 인식하는 단계;
인식된 복수의 바이트들 중에서 제 1 바이트 다음에 제 2 바이트가 연결되는, 바이트들 간의 고유 연결 구조의 출현 횟수를 계산하는 단계 - 상기 출현 횟수는 상기 인식된 복수의 바이트들 중 적어도 일부에 대해 계산됨 -;
및
상기 출현 횟수에 기초하여, 상기 입력 데이터를 상기 기 설정된 크기의 데이터로 변환하는 단계;
를 포함하는,
파일 특징 정보를 추출하는 방법.

청구항 7

제 6 항에 있어서,
상기 출현 횟수에 기초하여, 상기 입력 데이터를 상기 기 설정된 크기의 데이터로 변환하는 단계는,
상기 출현 횟수에 기초하여, 상기 제 1 바이트 다음에 상기 제 2 바이트가 출현되는, 바이트들 간의 연결 확률을 계산하는 단계; 및
상기 연결 확률에 기초하여, 상기 입력 데이터를 상기 기 설정된 크기의 데이터로 변환하는 단계;
를 포함하는,
파일 특징 정보를 추출하는 방법.

청구항 8

제 1 항에 있어서,
상기 기 설정된 크기의 데이터로부터 상기 입력 데이터의 특징 정보를 추출하는 단계는,

웨이블릿 변환(wavelet transform) 기법을 이용하여, 상기 기 설정된 크기의 데이터로부터 특징 정보를 추출하는 단계;

를 포함하는,

파일 특징 정보를 추출하는 방법.

청구항 9

제 8 항에 있어서,

상기 웨이블릿 변환(wavelet transform) 기법을 이용하여, 상기 기 설정된 크기의 데이터로부터 특징 정보를 추출하는 단계는,

상기 기 설정된 크기의 데이터에 상기 웨이블릿 변환을 수행하여, 웨이블릿 계수 행렬을 추출하는 단계 - 상기 웨이블릿 계수 행렬은 근사 계수 행렬(approximation coefficients matrix) 및 상세 계수 행렬(detail coefficients matrix)을 포함함 -; 및

상기 웨이블릿 계수 행렬에 기초하여, 상기 입력 데이터의 특징 정보를 추출하는 단계;

를 포함하는,

파일 특징 정보를 추출하는 방법.

청구항 10

제 9 항에 있어서,

상기 상세 계수 행렬은,

수평 계수 행렬(horizontal coefficients matrix), 수직 계수 행렬(vertical coefficients matrix) 또는 대각선 계수 행렬(diagonal coefficients matrix) 중 적어도 하나를 포함하는,

파일 특징 정보를 추출하는 방법.

청구항 11

제 10 항에 있어서,

상기 기 설정된 크기의 데이터는 행렬 구조를 갖는 데이터이고,

상기 특징 정보는,

상기 기 설정된 크기의 데이터가 1x1 행렬 구조가 되도록 제 1 횡수의 웨이블릿 변환을 수행하여 추출되는 상기 근사 계수 행렬과 관련된 제 1 특징, 제 1 임계 횡수 이상이고 그리고 상기 제 1 횡수 이하의 제 1 개수 범위 내에서의 제 2 횡수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬 값에 대한 제 2 특징, 제 1 임계 횡수 이상이고 그리고 상기 제 1 횡수 미만의 제 2 개수 범위 내에서의 제 3 횡수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬과 관련된 제 3 특징, 또는 상기 제 1 임계 횡수 미만의 제 2 임계 횡수 이상이고 그리고 상기 제 1 횡수 미만의 제 3 개수 범위 내에서의 제 4 횡수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬과 관련된 제 4 특징 중 적어도 하나를 포함하며, 여기서 상기 제 1 임계 횡수는 상기 제 1 횡수보다 작은,

파일 특징 정보를 추출하는 방법.

청구항 12

제 11 항에 있어서,

상기 제 3 특징은,

상기 제 3 횡수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬의 비율과 관련된 제 3-1 특징 또는 상기 제 3 횡수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬의 제 1 표준편차와 관련된 제 3-2 특징 중 적어도 하나를 포함하는,

파일 특징 정보를 추출하는 방법.

청구항 13

제 11 항에 있어서,

상기 제 4 특징은,

상기 제 4 횡수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬의 값의 합 관련된 제 4-1 특징, 상기 제 3 횡수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬의 값의 비대칭도와 관련된 제 4-2 특징, 상기 제 3 횡수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬의 제 2 표준편차와 관련된 제 4-3 특징 또는 상기 제 3 횡수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬의 평균 값과 관련된 제 4-4 특징 중 적어도 하나의 특징을 포함하는,

파일 특징 정보를 추출하는 방법.

청구항 14

제 1 항에 있어서,

바이너리 파일 형태의 입력 데이터로부터 악성 코드를 탐지하도록 상기 특징 정보를 학습 데이터로 이용하여 네트워크 모델을 학습시키는 단계;

를 더 포함하는,

파일 특징 정보를 추출하는 방법.

청구항 15

파일 특징 정보를 추출하는 컴퓨팅 장치로서,

바이너리(binary) 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환하는 바이너리 파일 변환부; 및

상기 기 설정된 크기의 데이터로부터 상기 입력 데이터의 특징 정보를 추출하는 신호처리부;

를 포함하는,

컴퓨팅 장치.

청구항 16

컴퓨터 판독가능 저장 매체에 저장되는 컴퓨터 프로그램으로서, 상기 컴퓨터 프로그램은 하나 이상의 프로세서에 의해 실행되는 경우 파일의 특징 정보를 추출하는 방법을 수행하며, 상기 방법은

바이너리(binary) 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환하는 단계; 및

상기 기 설정된 크기의 데이터로부터 상기 입력 데이터의 특징 정보를 추출하는 단계;

를 포함하는,

컴퓨터 판독가능 저장 매체에 저장되는 컴퓨터 프로그램.

발명의 설명

기술 분야

[0001] 본 개시는 악성코드 탐지에 관한 것으로, 구체적으로 바이너리 파일을 고정된 크기의 데이터로 변환한 뒤 웨이블릿 신호처리를 통해 정적으로 특징 정보를 추출하는 방법에 관한 것이다.

배경 기술

[0003] 전 세계적으로 대량의 악성코드가 사이버공간 상에 유포되고 있어 신속하게 바이너리 파일의 특징 정보를 추출하여 악성코드인지 여부를 탐지해야 할 필요성이 존재한다.

[0004] 그러나, 운영체제 환경에 따라 바이너리 파일은 PE(Portable Executable) 또는 ELF(Executable and Linkable Format) 등과 같은 구조를 가지고 있을 수 있다. 이에 따라, 바이너리 파일들에서 특징 정보를 추출하기 위해서는 해당 바이너리 파일의 구조에 따라 서로 다른 특징 정보 추출 과정이 필요할 수 있다. 또한, 패킹 및 난독화가 적용되지 않은 바이너리 파일과 패킹 및 난독화가 적용된 바이너리 파일에 대해서 특징 정보들을 추출하였을 때 패킹 및 난독화가 적용된 바이너리 파일에 대해 유의미한 특징 정보를 추출하기 어렵다는 한계도 존재할 수 있다. 더하여, 바이너리 파일에 대해 특징 정보를 추출하기 위해서는 바이너리 파일을 이미지화, 회색조, 배열화 등과 같은 데이터 변환과정을 수행해야 할 수 있다. 그러나, 바이너리 파일을 변환시킬 때 서로 다른 바이너리 파일 크기로 인하여 변환된 바이너리 파일의 형태가 고정되어 있지 않고, 이에 따라 서로 다른 크기의 바이너리 파일에서 균일한 수준의 특징 정보들을 추출하는데 한계가 존재할 수 있다. 더하여, 바이너리 파일에 대해 웨이블릿 신호처리를 통해 특징 정보를 추출할 경우 바이너리 파일의 크기에 따라 추출되는 특징 정보들이 동일하지 않다는 문제가 존재할 수 있다.

발명의 내용

해결하려는 과제

[0006] 본 개시는 전술한 배경기술에 대응하여 안출된 것으로, 신속하게 바이너리 파일의 특징 정보를 추출하고자 정적 특징 정보 추출 과정으로 추출한 특징 정보들을 사용하여 악성코드인지 탐지하고자 한다.

[0007] 또한, 운영체제 환경에 따라 바이너리 파일이 갖는 구조에 관계없이 모든 바이너리 파일에 대해 하나의 특징 정보 추출과정으로 특징 정보들을 추출할 수 있도록 한다.

[0008] 또한, 패킹 및 난독화가 적용되지 않은 바이너리 파일과 패킹 및 난독화가 적용된 바이너리 파일에서 서로 균일한 수준의 유의미한 바이너리 파일 특징 정보를 추출하고자 한다.

[0009] 또한, 바이너리 파일에 대해 padding, slicing, histogram, 히든 마르코프 등과 같은 방법들로 고정된 크기의 데이터를 생성하고자 한다.

[0010] 또한, 히든 마르코프 개념을 적용하여 바이너리 파일을 바이트 단위로 읽고 현재 바이트에서 다음 바이트로 연결되는 경우에 대한 확률값을 배열화하여 서로 다른 크기의 바이너리들을 고정된 형태의 데이터로 변환하고자 한다.

[0011] 또한, 고정된 형태의 데이터로 변환된 서로 다른 크기의 바이너리 파일들에서 웨이블릿 신호처리를 통하여 악성코드를 탐지하는데 유의미한 특징 정보들을 추출하고자 한다.

[0012] 본 개시의 기술적 과제들은 이상에서 언급한 기술적 과제로 제한되지 않으며, 언급되지 않은 또 다른 기술적 과제들은 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

- [0014] 전술한 바와 같은 과제를 해결하기 위한 본 개시의 일 실시예에 따라, 적어도 하나의 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 파일 특징 정보를 추출하는 방법이 개시된다. 상기 수행되는 파일 특징 정보를 추출하는 방법은, 바이너리(binary) 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환하는 단계; 및 상기 기 설정된 크기의 데이터로부터 상기 입력 데이터의 특징 정보를 추출하는 단계; 를 포함할 수 있다.
- [0015] 또한, 상기 바이너리(binary) 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환하는 단계는, 히든 마르코프 모델(hidden Markov model)을 이용하여, 상기 입력 데이터를 상기 기 설정된 크기의 데이터로 변환하는 단계; 를 포함할 수 있다.
- [0016] 또한, 상기 바이너리(binary) 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환하는 단계는, 상기 입력 데이터를 기억장치의 크기를 나타내는 제 1 단위(unit)로 구분하여 인식하는 단계; 상기 제 1 단위로 인식된 데이터의 개수를 카운팅하는 단계; 및 상기 제 1 단위의 데이터의 개수에 기초하여, 상기 입력 데이터를 상기 기 설정된 크기의 데이터로 변환하는 단계; 를 포함할 수 있다.
- [0017] 또한, 상기 바이너리(binary) 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환하는 단계는, 상기 입력 데이터를 바이트 단위로 인식하는 단계; 인식된 복수의 바이트들 각각의 개수를 카운팅하는 단계; 및 상기 복수의 바이트들 각각의 개수에 기초하여, 상기 입력 데이터를 상기 기 설정된 크기의 데이터로 변환하는 단계; 를 포함할 수 있다.
- [0018] 또한, 상기 복수의 바이트들 각각의 개수에 기초하여, 상기 입력 데이터를 상기 기 설정된 크기의 데이터로 변환하는 단계는, 상기 복수의 바이트들 각각의 개수에 기초하여, 전체 데이터 중 상기 복수의 바이트들 각각의 출현 확률을 계산하는 단계; 및 상기 출현 확률에 기초하여, 상기 입력 데이터를 상기 기 설정된 크기의 데이터로 변환하는 단계; 를 포함할 수 있다.
- [0019] 또한, 상기 바이너리(binary) 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환하는 단계는, 상기 입력 데이터를 바이트 단위로 인식하는 단계; 인식된 복수의 바이트들 중에서 제 1 바이트 다음에 제 2 바이트가 연결되는, 바이트들 간의 고유 연결 구조의 출현 횟수를 계산하는 단계 - 상기 출현 횟수는 상기 인식된 복수의 바이트들 중 적어도 일부에 대해 계산됨 -; 및 상기 출현 횟수에 기초하여, 상기 입력 데이터를 상기 기 설정된 크기의 데이터로 변환하는 단계; 를 포함할 수 있다.
- [0020] 또한, 상기 출현 횟수에 기초하여, 상기 입력 데이터를 상기 기 설정된 크기의 데이터로 변환하는 단계는, 상기 출현 횟수에 기초하여, 상기 제 1 바이트 다음에 상기 제 2 바이트가 출현되는, 바이트들 간의 연결 확률을 계산하는 단계; 및 상기 연결 확률에 기초하여, 상기 입력 데이터를 상기 기 설정된 크기의 데이터로 변환하는 단계; 를 포함할 수 있다.
- [0021] 또한, 상기 기 설정된 크기의 데이터로부터 상기 입력 데이터의 특징 정보를 추출하는 단계는, 웨이블릿 변환(wavelet transform) 기법을 이용하여, 상기 기 설정된 크기의 데이터로부터 특징 정보를 추출하는 단계; 를 포함할 수 있다.
- [0022] 또한, 상기 웨이블릿 변환(wavelet transform) 기법을 이용하여, 상기 기 설정된 크기의 데이터로부터 특징 정보를 추출하는 단계는, 상기 기 설정된 크기의 데이터에 상기 웨이블릿 변환을 수행하여, 웨이블릿 계수 행렬을 추출하는 단계 - 상기 웨이블릿 계수 행렬은 근사 계수 행렬(approximation coefficients matrix) 및 상세 계수 행렬(detail coefficients matrix)을 포함함 -; 및 상기 웨이블릿 계수 행렬에 기초하여, 상기 입력 데이터의 특징 정보를 추출하는 단계; 를 포함할 수 있다.
- [0023] 또한, 상기 상세 계수 행렬은, 수평 계수 행렬(horizontal coefficients matrix), 수직 계수 행렬(vertical coefficients matrix) 또는 대각선 계수 행렬(diagonal coefficients matrix) 중 적어도 하나를 포함할 수 있다.
- [0024] 또한, 상기 기 설정된 크기의 데이터는 행렬 구조를 갖는 데이터이고, 상기 특징 정보는, 상기 기 설정된 크기의 데이터가 1x1 행렬 구조가 되도록 제 1 횟수의 웨이블릿 변환을 수행하여 추출되는 상기 근사 계수 행렬과 관련된 제 1 특징, 제 1 임계 횟수 이상이고 그리고 상기 제 1 횟수 이하의 제 1 개수 범위 내에서의 제 2 횟수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬 값에 대한 제 2 특징, 상기 제 1 임계 횟수 이상이고 그리고 상기 제 1 횟수 미만의 제 2 개수 범위 내에서의 제 3 횟수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬과 관련된 제 3 특징, 또는 상기 제 1 임계 횟수 미만의 제 2 임계 횟수 이상이고 그리고

상기 제 1 횡수 미만의 제 3 개수 범위 내에서의 제 4 횡수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬과 관련된 제 4 특징 중 적어도 하나를 포함하며, 여기서 상기 제 1 임계 횡수는 상기 제 1 횡수보다 작을 수 있다.

[0025] 또한, 상기 제 3 특징은, 상기 제 3 횡수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬의 비율과 관련된 제 3-1 특징 또는 상기 제 3 횡수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬의 제 1 표준편차와 관련된 제 3-2 특징 중 적어도 하나를 포함할 수 있다.

[0026] 또한, 상기 제 4 특징은, 상기 제 4 횡수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬의 값의 합과 관련된 제 4-1 특징, 상기 제 3 횡수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬의 값의 비대칭도와 관련된 제 4-2 특징, 상기 제 3 횡수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬의 제 2 표준편차와 관련된 제 4-3 특징 또는 상기 제 3 횡수의 웨이블릿 변환을 수행하여 추출되는 상기 대각선 계수 행렬의 평균 값과 관련된 제 4-4 특징 중 적어도 하나의 특징을 포함할 수 있다.

[0027] 또한, 바이너리 파일 형태의 입력 데이터로부터 악성 코드를 탐지하도록 상기 특징 정보를 학습 데이터로 이용하여 네트워크 모델을 학습시키는 단계;를 더 포함할 수 있다.

[0028] 또한, 파일 특징 정보를 추출하는 컴퓨팅 장치로서, 바이너리(binary) 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환하는 바이너리 파일 변환부; 및 상기 기 설정된 크기의 데이터로부터 상기 입력 데이터의 특징 정보를 추출하는 신호처리부;를 포함할 수 있다.

[0029] 또한, 컴퓨터 판독가능 저장 매체에 저장되는 컴퓨터 프로그램으로서, 상기 컴퓨터 프로그램은 하나 이상의 프로세서에 의해 실행되는 경우 파일의 특징 정보를 추출하는 방법을 수행하며, 상기 방법은 바이너리(binary) 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환하는 단계; 및 상기 기 설정된 크기의 데이터로부터 상기 입력 데이터의 특징 정보를 추출하는 단계;를 포함할 수 있다.

[0031] 본 개시에서 얻을 수 있는 기술적 해결 수단은 이상에서 언급한 해결 수단들로 제한되지 않으며, 언급하지 않은 또 다른 해결 수단들은 아래의 기재로부터 본 개시가 속하는 기술분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

발명의 효과

[0033] 정적으로 특징 정보들을 추출하여 신속하게 바이너리 파일이 악성코드인지 탐지할 수 있으며 바이너리 파일을 실행시키지 않아 악성코드 감염에 대한 위험을 제거할 수 있다.

[0034] 또한, 운영체제 환경에 관계없이 모든 바이너리 파일들에 대해 특징 정보들을 추출할 수 있다.

[0035] 또한, 바이너리 파일이 패킹 및 난독화 되어있는지 여부에 관계없이 균일한 수준의 특징 정보들을 추출할 수 있다.

[0036] 또한, 바이너리 파일에서 특징 정보들을 추출할 때 입력되는 바이너리 파일을 본 발명에서의 데이터 변환과정을 통해 바이너리 파일의 서로 다른 크기에 관계없이 고정된 형태의 데이터로 변환하여 균일한 수준의 특징 정보들을 추출할 수 있다.

[0037] 또한, 데이터 변환과정으로 고정된 형태의 데이터에서 웨이블릿 신호처리를 통해 악성코드 탐지에 유의미한 특징 정보들을 추출할 수 있다.

[0038] 본 개시에서 얻을 수 있는 효과는 이상에서 언급한 효과로 제한되지 않으며, 언급하지 않은 또 다른 효과들은 아래의 기재로부터 본 개시가 속하는 기술분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0040] 다양한 양상들이 이제 도면들을 참조로 기재되며, 여기서 유사한 참조 번호들은 총괄적으로 유사한 구성요소들

을 지칭하는데 이용된다. 이하의 실시예에서, 설명 목적을 위해, 다수의 특정 세부사항들이 하나 이상의 양상들의 총체적 이해를 제공하기 위해 제시된다. 그러나, 그러한 양상(들)이 이러한 특정 세부사항들 없이 실시될 수 있음은 명백할 것이다. 다른 예시들에서, 공지의 구조들 및 장치들이 하나 이상의 양상들의 기재를 용이하게 하기 위해 블록도 형태로 도시된다.

도 1은 본 개시의 몇몇 실시예에 따른 컴퓨팅 장치의 일례를 설명하기 위한 블록 구성도이다.

도 2는 본 개시의 몇몇 실시예에 따른 컴퓨팅 장치가 입력 데이터의 특징 정보를 추출하는 방법의 일례를 설명하기 위한 흐름도이다.

도 3은 본 개시의 몇몇 실시예에 따른 컴퓨팅 장치가 바이트들 각각의 개수에 기초하여 입력 데이터를 기 설정된 크기의 데이터로 변환하는 방법의 일례를 설명하기 위한 흐름도이다.

도 4는 본 개시의 몇몇 실시예에 따른 컴퓨팅 장치가 바이트들 각각의 출현 확률에 기초하여 입력 데이터를 기 설정된 크기의 데이터로 변환하는 방법의 일례를 설명하기 위한 흐름도이다.

도 5는 본 개시의 몇몇 실시예에 따른 컴퓨팅 장치가 바이트들 간의 고유 연결 구조의 출현 횟수에 기초하여 입력 데이터를 기 설정된 크기의 데이터로 변환하는 방법의 일례를 설명하기 위한 흐름도이다.

도 6은 본 개시의 몇몇 실시예에 따른 컴퓨팅 장치가 바이트들 간의 연결 확률에 기초하여 입력 데이터를 기 설정된 크기의 데이터로 변환하는 방법의 일례를 설명하기 위한 흐름도이다.

도 7은 컴퓨팅 장치가 웨이블릿 변환 기법을 이용하여 기 설정된 크기의 데이터로부터 특징 정보를 추출하는 방법의 일례를 설명하기 위한 흐름도이다.

도 8은 본 개시내용의 실시예들이 구현될 수 있는 예시적인 컴퓨팅 환경에 대한 일반적인 개략도를 도시한다.

발명을 실시하기 위한 구체적인 내용

- [0041] 다양한 실시예들 및/또는 양상들이 이제 도면들을 참조하여 개시된다. 하기 설명에서는 설명을 목적으로, 하나 이상의 양상들의 전반적 이해를 돕기 위해 다수의 구체적인 세부사항들이 개시된다. 그러나, 이러한 양상(들)은 이러한 구체적인 세부사항들 없이도 실행될 수 있다는 점 또한 본 개시의 기술 분야에서 통상의 지식을 가진 자에게 감지될 수 있을 것이다. 이후의 기재 및 첨부된 도면들은 하나 이상의 양상들의 특정한 예시적인 양상들을 상세하게 기술한다. 하지만, 이러한 양상들은 예시적인 것이고 다양한 양상들의 원리들에서의 다양한 방법들 중 일부가 이용될 수 있으며, 기술되는 설명들은 그러한 양상들 및 그들의 균등물들을 모두 포함하고자 하는 의도이다. 구체적으로, 본 명세서에서 사용되는 "실시예", "예", "양상", "예시" 등은 기술되는 임의의 양상 또는 설계가 다른 양상 또는 설계들보다 양호하다거나, 이점이 있는 것으로 해석되지 않을 수도 있다.
- [0042] 이하, 도면 부호에 관계없이 동일하거나 유사한 구성 요소는 동일한 참조 번호를 부여하고 이에 대한 중복되는 설명은 생략한다. 또한, 본 명세서에 개시된 실시예를 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 본 명세서에 개시된 실시예의 요지를 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다. 또한, 첨부된 도면은 본 명세서에 개시된 실시예를 쉽게 이해할 수 있도록 하기 위한 것일 뿐, 첨부된 도면에 의해 본 명세서에 개시된 기술적 사상이 제한되지 않는다.
- [0043] 비록 제 1, 제 2 등이 다양한 소자나 구성요소들을 서술하기 위해서 사용되나, 이들 소자나 구성요소들은 이들 용어에 의해 제한되지 않음은 물론이다. 이들 용어들은 단지 하나의 소자나 구성요소를 다른 소자나 구성요소와 구별하기 위하여 사용하는 것이다. 따라서, 이하에서 언급되는 제 1 소자나 구성요소는 본 발명의 기술적 사상 내에서 제 2 소자나 구성요소 일 수도 있음은 물론이다.
- [0044] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다.
- [0045] 더불어, 용어 "또는"은 배타적 "또는"이 아니라 내포적 "또는"을 의미하는 것으로 의도된다. 즉, 달리 특정되지 않거나 문맥상 명확하지 않은 경우에, "X는 A 또는 B를 이용한다"는 자연적인 내포적 치환 중 하나를 의미하는 것으로 의도된다. 즉, X가 A를 이용하거나; X가 B를 이용하거나; 또는 X가 A 및 B 모두를 이용하는 경우, "X는 A 또는 B를 이용한다"가 이들 경우들 어느 것으로도 적용될 수 있다. 또한, 본 명세서에 사용된 "및/또는"이라는 용어는 열거된 관련 아이템들 중 하나 이상의 아이템의 가능한 모든 조합을 지칭하고 포함하는 것으로 이해

되어야 한다.

- [0046] 또한, "포함한다" 및/또는 "포함하는"이라는 용어는, 해당 특징 및/또는 구성요소가 존재함을 의미하지만, 하나 이상의 다른 특징, 구성요소 및/또는 이들의 그룹의 존재 또는 추가를 배제하지 않는 것으로 이해되어야 한다. 또한, 달리 특정되지 않거나 단수 형태를 지시하는 것으로 문맥상 명확하지 않은 경우에, 본 명세서와 청구범위에서 단수는 일반적으로 "하나 또는 그 이상"을 의미하는 것으로 해석되어야 한다.
- [0047] 더불어, 본 명세서에서 사용되는 용어 "정보" 및 "데이터"는 종종 서로 상호교환 가능하도록 사용될 수 있다.
- [0048] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성 요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성 요소가 존재하지 않는 것으로 이해되어야 할 것이다.
- [0049] 이하의 설명에서 사용되는 구성 요소에 대한 접미사 "모듈" 및 "부"는 명세서 작성의 용이함만이 고려되어 부여되거나 혼용되는 것으로서 그 자체로 서로 구별되는 의미 또는 역할을 갖는 것은 아니다.
- [0050] 본 개시의 목적 및 효과, 그리고 그것들을 달성하기 위한 기술적 구성들은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 본 개시를 설명하는데 있어서 공지 기능 또는 구성에 대한 구체적인 설명이 본 개시의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이다. 그리고 후술되는 용어들은 본 개시에서의 기능을 고려하여 정의된 용어들로써 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다.
- [0051] 그러나 본 개시는 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있다. 단지 본 실시예들은 본 개시가 완전하도록 하고, 본 개시가 속하는 기술분야에서 통상의 지식을 가진 자에게 개시의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 개시는 청구항의 범주에 의해 정의될 뿐이다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.
- [0053] 본 개시에서, 컴퓨팅 장치는 바이너리(binary) 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환할 수 있다. 바이너리 파일은 데이터의 저장과 처리를 목적으로 인코딩된 파일로서, 운영체제 환경에 따라 상이한 구조를 가질 수 있다. 일례로, 윈도우 계열의 운영 체제에서의 바이너리 파일은 PE(Portable Executable) 포맷일 수 있고, 유닉스 계열의 운영 체제에서의 바이너리 파일은 ELF 포맷(Executable and Linkable Format)일 수 있다. 종래에는 운영체제 환경 별로 상이한 프로세스를 통해 바이너리 파일로부터 특징 정보를 추출할 수 있었다. 이 경우 추출된 특징 정보들이 균일하지 못하다는 문제점이 존재할 수 있다. 본 개시에서는 바이너리(binary) 파일 형태의 입력 데이터에 전처리를 수행하여 기 설정된 크기의 데이터로 변환할 수 있다. 본 개시에서 수행되는 전처리는 운영 체제 환경과는 상관없이 동일한 전처리일 수 있다. 따라서, 변환된 기 설정된 크기의 데이터로부터 입력 데이터의 특징 정보를 추출하는 경우, 추출된 특징 정보들의 균일성이 담보될 수 있다. 더하여, 컴퓨팅 장치는 바이너리 파일을 실행시키지 않고, 정적으로 특징 정보를 추출할 수 있다. 따라서, 악성 코드 감염에 대한 위험도 제거될 수 있다. 이하, 도 1 내지 도 7을 통해 컴퓨팅 장치가 파일 특징 정보를 추출하는 방법에 대해 설명한다.
- [0055] 도 1은 본 개시의 몇몇 실시예에 따른 컴퓨팅 장치의 일례를 설명하기 위한 블록 구성도이다.
- [0056] 도 1을 참조하면, 컴퓨팅 장치(100)는 프로세서(110), 바이너리 파일 변환부(120) 및 신호처리부(130)를 포함할 수 있다. 다만, 상술한 구성 요소들은 컴퓨팅 장치(100)를 구현하는데 있어서 필수적인 것은 아니어서, 컴퓨팅 장치(100)는 위에서 열거된 구성요소들 보다 많거나, 또는 적은 구성요소들을 가질 수 있다.
- [0057] 컴퓨팅 장치(100)는 예를 들어, 마이크로프로세서, 메인프레임 컴퓨터, 디지털 프로세서, 휴대용 디바이스 또는 디바이스 제어기 등과 같은 임의의 타입의 컴퓨터 시스템 또는 컴퓨터 디바이스를 포함할 수 있다.
- [0058] 프로세서(110)는 통상적으로 컴퓨팅 장치(100)의 전반적인 동작을 처리할 수 있다. 프로세서(110)는 컴퓨팅 장치(100)에 포함된 구성요소들을 통해 입력 또는 출력되는 신호, 데이터, 정보 등을 처리하거나 저장부에 저장된 응용 프로그램을 구동함으로써, 사용자에게 적절한 정보 또는 기능을 제공 또는 처리할 수 있다.
- [0059] 바이너리(binary) 파일 변환부(120)는 바이너리 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환할

수 있다. 기 설정된 크기의 데이터는 행렬 구조의 데이터일 수 있다. 일례로, 기 설정된 크기의 데이터는 256x256 크기를 갖는 행렬 구조의 데이터일 수 있다. 다만, 기 설정된 크기의 데이터가 이에 한정되는 것은 아니며, 다양한 크기를 갖는 행렬 구조의 데이터일 수 있다. 일례로, 기 설정된 크기의 데이터는 512x512 크기를 갖는 행렬 구조의 데이터일 수도 있고, 128x128 크기를 갖는 행렬 구조의 데이터일 수도 있다.

[0060] 본 개시의 몇몇 실시예에 따르면, 바이너리 파일 변환부(120)는 입력 데이터를 기 설정된 크기의 데이터로 변환하기 위해, 입력 데이터를 기억장치의 크기를 나타내는 제 1 단위(unit)로 구분하여 인식할 수 있다. 일 실시예에서, 제 1 단위는 컴퓨팅 장치의 저장 단위를 포함할 수 있다. 제 1 단위는 일례로, 바이트(byte), 페이지, 블록 또는 비트(bit) 등의 단위일 수 있다. 바이너리 파일 변환부(120)는 제 1 단위로 인식된 데이터의 개수를 카운팅할 수 있다. 바이너리 파일 변환부(120)는 제 1 단위로 인식된 데이터의 개수에 기초하여, 입력 데이터를 기 설정된 크기의 데이터로 변환할 수 있다.

[0061] 본 개시의 몇몇 실시예에 따르면, 바이너리 파일 변환부(120)는 바이너리 파일 관독부, 바이트 확률 계산부 및 고정 데이터 생성부를 포함할 수 있다.

[0062] 바이너리 파일 관독부는 바이너리 파일 형태의 입력 데이터를 제 1 단위로 구분하여 인식할 수 있다. 일례로, 바이너리 파일 관독부는 바이너리 파일 형태의 입력 데이터를 바이트 단위로 구분하여 인식할 수 있다.

[0063] 바이트 확률 계산부는 바이너리 파일 관독부를 통해 입력 데이터가 바이트 단위로 인식된 경우, 인식된 복수의 바이트들 각각의 개수를 카운팅할 수 있다. 바이트 확률 계산부는 복수의 바이트들 각각의 개수에 기초하여, 전체 데이터 중 복수의 바이트들 각각의 출현 확률을 계산할 수 있다. 출현 확률은 전체 데이터 중 복수의 바이트들 각각이 얼마나 출현했는지를 나타내는 확률일 수 있다. 일례로, 복수의 바이트들 중 제 1 바이트에 대해 계산되는 제 1 출현 확률은, 복수의 바이트들 중 제 1 바이트가 차지하는 비율을 나타낼 수는 값일 수도 있다.

[0064] 고정 데이터 생성부는 바이트 확률 계산부를 통해 계산된 출현 확률에 기초하여, 입력 데이터를 기 설정된 크기의 데이터로 변환할 수 있다. 이하, 바이너리 파일 변환부(120)가 바이너리 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환하는 방법에 대한 일례는 도 2 내지 도 6을 통해 설명한다.

[0065] 한편, 신호처리부(130)는 기 설정된 크기의 데이터로부터 입력 데이터의 특징 정보를 추출할 수 있다. 특징 정보는 바이너리 파일 형태의 입력 데이터가 악성코드인지 여부를 탐지하기 위해 사용되는 정보일 수 있다.

[0066] 일례로, 프로세서(110)는 추출된 특징 정보를 학습 데이터로 이용하여, 바이너리 파일 형태의 입력 데이터로부터 악성 코드를 탐지하도록 네트워크 모델을 학습시킬 수 있다. 본 개시에 따른 방법에 따라 추출된 특징 정보를 학습 데이터로 이용하여 네트워크 모델을 학습시키는 경우, 학습된 네트워크 모델은 높은 정확도로 악성 코드를 탐지할 수 있음은 하기의 표 1 내지 표 3과 같이 입증될 수 있다.

표 1

	precision	recall	F1-score	Support
0	0.9214	0.8845	0.9026	1,671
1	0.9552	0.9703	0.9627	4,237
Macro avg	0.9383	0.9274	0.9326	5,908
Weighted avg	0.9456	0.9460	0.9457	5,908

표 2

	precision	recall	F1-score	Support
0	0.9354	0.8934	0.9139	2,870
1	0.9573	0.9749	0.9660	7,041
Macro avg	0.9464	0.9341	0.9400	9,911
Weighted avg	0.9510	0.9513	0.9509	9,911

표 3

	precision	recall	F1-score	Support
0	0.9577	0.9589	0.9583	3,214

1	0.9702	0.9693	0.9697	4,427
Macro avg	0.9639	0.9641	0.9640	7,641
Weighted avg	0.9649	0.9649	0.9649	7,641

[0070] 표 1은 KISA 2017 악성코드 데이터에 대한 성능 평가 결과일 수 있다. 표 2는 KISA 2018 악성코드 데이터에 대한 성능 평가 결과일 수 있다. 표 3은 KISA 2019 악성코드 데이터에 대한 성능 평가 결과일 수 있다. KISA 악성코드는 한국인터넷진흥원(Korea Internet & Security Agency)에서 제공하는 데이터일 수 있다.

[0071] 표 1 내지 표 3의 0은 비악성(benign) 코드이고, 1은 악성(malicious) 코드일 수 있다. Macro AVG는 TP(True Positive), FN(False Negative), FP(False Positive) 및 TN(True Negative) 각각을 계산한 후 합친 값일 수 있다. Weighted AVG는 각 클래스에 해당하는 데이터의 개수에 기 설정된 가중치를 주어 평균을 구한 값일 수 있다.

[0072] 표 1을 참조하면, 정확도(accuracy)가 0.9460으로 나타나고, AUC(Area Under the curve) Score가 0.9274로 나타나는 것을 확인할 수 있다. 표 2를 참조하면, 정확도가 0.9513으로 나타나고, AUC Score가 0.9341로 나타나는 것을 확인할 수 있다. 표 3을 참조하면, 정확도가 0.9649으로 나타나고, AUC Score가 0.9641로 나타나는 것을 확인할 수 있다.

[0073] 정확도는 하기의 수학적 식 1과 같이 계산되는 값일 수 있으며, 정확도가 높을수록 모델의 성능이 높은 것으로 평가될 수 있다.

수학적 식 1

$$\frac{TP + TN}{TP + FP + TN + FN}$$

[0074]

[0075] AUC Score는 ROC(Receiver Operating Characteristic) 커브의 아래쪽 면적에 해당하는 값으로서, AUC Score가 높을수록 모델이 악성 코드를 우수한 성능으로 예측했다는 것을 의미할 수 있다.

[0076] 상기 표 1 내지 표 3과 같이 본 개시에 따른 방법에 따라 추출된 특징 정보를 학습 데이터로 이용하여 네트워크 모델을 학습시키는 경우, 학습된 네트워크 모델은 높은 정확도로 악성 코드를 탐지 및 예측할 수 있다.

[0077] 한편, 본 개시에서 특징 정보를 학습데이터로 이용하여 학습되는 네트워크 모델은 컨볼루션 뉴럴 네트워크(Convolutional Neural Networks, CNN), ANN(Artificial Neural Network), DNN(Deep Neural Network) 또는 RNN(Recurrent Neural Network) 등일 수 있다.

[0078] 이하에서는 본 개시에 따른 컴퓨팅 장치(100)가 입력 데이터의 특징 정보를 추출하는 방법에 대해 설명한다.

[0080] 도 2는 본 개시의 몇몇 실시예에 따른 컴퓨팅 장치가 입력 데이터의 특징 정보를 추출하는 방법의 일례를 설명하기 위한 흐름도이다.

[0081] 도 2를 참조하면, 컴퓨팅 장치(100)의 바이너리 파일 변환부(120)는 바이너리 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환할 수 있다(S110).

[0082] 실시예에 따라, 바이너리 파일 변환부(120)는 입력 데이터를 기억장치의 크기를 나타내는 제 1 단위로 구분하여 인식할 수 있다. 바이너리 파일 변환부(120)는 인식된 제 1 단위 데이터의 개수를 카운팅하고, 제 1 단위의 데이터의 개수에 기초하여, 입력 데이터를 기 설정된 크기의 데이터로 변환할 수 있다.

[0083] 본 개시의 몇몇 실시예에 따르면, 바이너리 파일 변환부(120)는 히든 마르코프 모델(hidden Markov model)을 이용하여, 입력 데이터를 기 설정된 크기의 데이터로 변환할 수 있다. 히든 마르코프 모델은 통계 기반의 모델로서, 입력 데이터가 바이트 단위로 인식된 경우, 바이트들 간의 연결 확률에 기초하여 입력 데이터를 변환할 수 있는 모델일 수 있다. 연결 확률은 복수의 바이트들 중 제 1 바이트 다음에 제 2 바이트가 출현하여, 연결되는 확률일 수 있다. 실시예에 따라, 바이너리 파일 변환부(120)는 padding, slicing 또는 histogram과 같은 방법들을 이용하여 입력 데이터를 기 설정된 크기의 데이터로 변환할 수도 있다. 이하, 바이너리 파일 변환부(120)가

히든 마르코프 모델을 이용하여, 입력 데이터를 기 설정된 크기의 데이터로 변환하는 방법의 일례는 도 5 및 도 6을 통해 설명한다.

[0084] 신호처리부(130)는 기 설정된 크기의 데이터로부터 입력 데이터의 특징 정보를 추출할 수 있다(S120).

[0085] 구체적으로, 신호처리부(130)는 웨이블릿 변환(wavelet transform) 기법을 이용하여, 기 설정된 크기의 데이터로부터 특징 정보를 추출할 수 있다. 웨이블릿 변환 기법은 오디오 신호, 이미지 또는 동영상 등 다양한 종류의 데이터로부터 정보를 추출하는데 사용될 수 있는 시그널 프로세싱 기법으로서, 시간-주파수의 해석이 가능한 기법일 수 있다. 웨이블릿 변환은 저주파수 정보의 분석을 위해서는 장시간의 간격을 이용하고, 고주파수 정보를 위해서는 단시간의 간격을 이용하여 시간적 또는 공간적 추이를 동시에 가변적으로 다룰 수 있다. 또 다른 시그널 프로세싱 기법 중 하나인 푸리에 변환(Fourier transform) 기법도 시간-주파수의 해석이 가능할 수 있으며, 시간이나 공간에 대한 함수를 시간 정보 또는 공간 주파수 정보로 변환할 수 있다. 그러나, 푸리에 변환 기법의 경우, 사용자가 시간 축을 어떻게 설정하느냐에 따라, 시간 정보 또는 공간 주파수 정보가 변화될 수 있다. 일례로, 푸리에 기법은 시간 정보가 많이 추출될 수 있도록 사용자가 시간 축을 설정하는 경우, 공간 정보에서 손해가 나타날 수 있다. 반대로, 푸리에 기법은 공간 정보가 많이 추출될 수 있도록 사용자가 시간 축을 설정하는 경우, 시간 정보에서 손해가 나타날 수 있다. 반면, 웨이블릿 변환은 시간 정보 및 공간 정보를 균등하게 추출할 수 있으며, 다시 말해 상술한 푸리에 변환의 단점을 해결할 수 있다. 따라서, 본 개시에서는 웨이블릿 변환 기법을 이용하여, 기 설정된 크기의 데이터로부터 특징 정보를 추출할 수 있다. 이하, 신호처리부(130)가 웨이블릿 변환 기법을 이용하여, 기 설정된 크기의 데이터로부터 특징 정보를 추출하는 방법에 대한 일례는 도 7을 통해 설명한다.

[0086] 상술한 구성에 따르면, 컴퓨팅 장치(100)는 운영체제 환경에 따라 상이한 바이너리 파일 형태의 입력 데이터를 모두 기 설정된 크기의 데이터로 변환할 수 있다. 그리고, 컴퓨팅 장치는 기 설정된 크기의 데이터로부터 입력 데이터의 특징 정보를 추출할 수 있다. 이에 따라, 컴퓨팅 장치(100)는 상이한 운영체제 환경에서의 데이터가 입력되더라도, 서로 균일한 수준의 유의미한 특징 정보를 추출할 수 있다.

[0087] 이하에서는 컴퓨팅 장치(100)가 입력 데이터를 기 설정된 크기의 데이터로 변환하는 일례에 대해 좀 더 설명한다.

[0089] 도 3은 본 개시의 몇몇 실시예에 따른 컴퓨팅 장치가 바이트들 각각의 개수에 기초하여 입력 데이터를 기 설정된 크기의 데이터로 변환하는 방법의 일례를 설명하기 위한 흐름도이다. 도 4는 본 개시의 몇몇 실시예에 따른 컴퓨팅 장치가 바이트들 각각의 출현 확률에 기초하여 입력 데이터를 기 설정된 크기의 데이터로 변환하는 방법의 일례를 설명하기 위한 흐름도이다.

[0090] 도 3을 참조하면, 컴퓨팅 장치(100)의 바이너리 파일 변환부(120)는 입력 데이터를 바이트 단위로 인식할 수 있다(S111).

[0091] 예를 들어, 바이너리 파일 형태의 입력 데이터는 0~255 사이의 값들을 포함할 수 있다. 바이너리 파일 변환부(120)는 0~255 사이의 값들 각각을 바이트 단위로 인식할 수 있다. 일례로, 바이너리 파일 변환부(120)는 0~255 사이의 값들을 0, 1, 2, 3 등과 같이 바이트 단위로 인식할 수 있다.

[0092] 바이너리 파일 변환부(120)는 인식된 복수의 바이트들 각각의 개수를 카운팅할 수 있다(S112).

[0093] 예를 들어, 바이너리 파일 변환부(120)는 0인 값이 몇 개인지, 1인 값이 몇 개인지 등과 같이 복수의 바이트들 각각의 개수를 카운팅할 수 있다.

[0094] 바이너리 파일 변환부(120)는 복수의 바이트들 각각의 개수에 기초하여, 입력 데이터를 기 설정된 크기의 데이터로 변환할 수 있다(S113).

[0095] 예를 들어, 바이너리 파일 변환부(120)는 입력 데이터를 복수의 바이트들 각각의 개수에 대응하는 크기를 갖는 데이터로 변환할 수 있다.

[0096] 한편, 도 4를 참조하면, 바이너리 파일 변환부(120)는 복수의 바이트들 각각의 개수가 카운팅된 경우, 복수의 바이트들 각각의 개수에 기초하여 전체 데이터 중 복수의 바이트들 각각의 출현 확률을 계산할 수 있다(S1131). 출현 확률은 전체 데이터 중 복수의 바이트들 각각이 얼마나 출현했는지를 나타내는 확률일 수 있다. 일례로, 복수의 바이트들 중 제 1 바이트에 대해 계산되는 제 1 출현 확률은, 복수의 바이트들 중 제 1 바이트가 차지하

는 비율을 나타낼 수는 값일 수도 있다.

- [0097] 예를 들어, 바이너리 파일 변환부(120)는 0~255 사이의 값들 중 0이 몇 퍼센트 출현하는지, 1이 몇 퍼센트 출현하는지 또는 2가 몇 퍼센트 출현하는지 등을 계산할 수 있다.
- [0098] 바이너리 파일 변환부(120)는 출현 확률에 기초하여, 입력 데이터를 기 설정된 크기의 데이터로 변환할 수 있다 (1132).
- [0099] 예를 들어, 바이너리 파일 변환부(120)는 입력 데이터를 복수의 바이트들 각각의 출현 확률에 대응하는 크기를 갖는 데이터로 변환할 수 있다.
- [0100] 상술한 구성에 따르면, 컴퓨팅 장치(100)는 운영체제 환경에 따라 상이한 바이너리 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환할 수 있다. 다시 말해, 컴퓨팅 장치(100)는 운영체제 환경에 따라 상이한 바이너리 파일 형태의 입력 데이터가 입력되더라도, 입력 데이터를 모두 균일한 크기의 데이터로 변환할 수 있다. 이에 따라, 컴퓨팅 장치(100)가 추출하는 입력 데이터의 특징 정보도 균일할 수 있다.
- [0101] 이하에서는 컴퓨팅 장치(100)가 입력 데이터를 기 설정된 크기의 데이터로 변환하는 다른 일례에 대해 설명한다.
- [0103] 도 5는 본 개시의 몇몇 실시예에 따른 컴퓨팅 장치가 바이트들 간의 고유 연결 구조의 출현 횟수에 기초하여 입력 데이터를 기 설정된 크기의 데이터로 변환하는 방법의 일례를 설명하기 위한 흐름도이다. 도 6은 본 개시의 몇몇 실시예에 따른 컴퓨팅 장치가 바이트들 간의 연결 확률에 기초하여 입력 데이터를 기 설정된 크기의 데이터로 변환하는 방법의 일례를 설명하기 위한 흐름도이다.
- [0104] 본 개시의 몇몇 실시예에 따르면, 컴퓨팅 장치(100)의 바이너리 파일 변환부(120)는 히든 마르코프 모델을 이용하여, 입력 데이터를 기 설정된 크기의 데이터로 변환할 수 있다. 히든 마르코프 모델은 입력 데이터에 포함된 바이트들 간의 연결 확률에 기초하여 입력 데이터를 변환할 수 있는 모델일 수 있다. 연결 확률은 복수의 바이트들 중 제 1 바이트 다음에 제 2 바이트가 출현하여, 연결되는 확률일 수 있다.
- [0105] 구체적으로, 도 5를 참조하면, 바이너리 파일 변환부(120)는 입력 데이터를 바이트 단위로 인식할 수 있다 (S210).
- [0106] 바이너리 파일 변환부(120)는 인식된 복수의 바이트들 중에서 제 1 바이트 다음에 제 2 바이트가 연결되는, 바이트들 간의 고유 연결 구조의 출현 횟수를 계산할 수 있다(S220). 여기서, 제 1 바이트 및 제 2 바이트는 설명의 편의를 위하여 구분한 것으로서, 특정 바이트로 한정되는 것은 아니다. 일례로, 바이너리 파일 변환부(120)는 인식된 복수의 바이트들 중 적어도 일부에 대해 출현 횟수를 계산할 수 있다.
- [0107] 고유 연결 구조는 적어도 두개의 바이트들 간의 고유한 연결 구조로서, 다른 적어도 두개의 바이트들 간의 연결 구조와는 구분되는 되는 것으로 이해될 수 있다.
- [0108] 예를 들어, 바이너리 파일 변환부(120)는 0의 값을 갖는 제 1 바이트 다음에 1의 값을 갖는 제 2 바이트가 연결되는 구조를 고유 연결 구조라고 결정할 수 있다. 예를 들어, 바이너리 파일 변환부(120)는 0의 값을 갖는 제 1 바이트 다음에 1의 값을 갖는 제 2 바이트가 연결되는 구조를 고유 연결 구조라고 결정할 수 있다. 다른 예를 들어, 바이너리 파일 변환부(120)는 1의 값을 갖는 제 2 바이트 다음에 2의 값을 갖는 제 3 바이트가 연결되는 구조를 고유 연결 구조라고 결정할 수 있다.
- [0109] 출현 횟수는 고유 연결 구조가 인식된 복수의 바이트들 내에서 출현한 횟수를 나타낼 수 있다.
- [0110] 예를 들어, 바이너리 파일 변환부(120)는 0의 값을 갖는 제 1 바이트 다음에 1의 값을 갖는 제 2 바이트가 연결되는 고유 연결 구조의 출현 횟수를 계산할 수 있다. 다른 예를 들어, 바이너리 파일 변환부(120)는 1의 값을 갖는 제 2 바이트 다음에 2의 값을 갖는 제 3 바이트가 연결되는 고유 연결 구조의 출현 횟수를 계산할 수 있다.
- [0111] 바이너리 파일 변환부(120)는 바이트들 간의 고유 연결 구조의 출현 횟수에 기초하여, 입력 데이터를 기 설정된 크기의 데이터로 변환할 수 있다(S230).
- [0112] 실시예에 따라, 바이너리 파일 변환부(120)는 입력 데이터를 256x256 크기를 갖는 행렬 구조의 데이터로 변환할 수 있다. 다만, 기 설정된 크기는 256x256 크기로 한정되는 것은 아니며, 이보다 더 크거나 또는 더 작은 크기

일 수도 있다.

- [0113] 한편, 도 6을 참조하면, 바이너리 파일 변환부(120)는 바이트들 간의 고유 연결 구조의 출현 횟수에 기초하여, 제 1 바이트 다음에 제 2 바이트가 출현되는, 바이트들 간의 연결 확률을 계산할 수 있다(S231). 연결 확률은 복수의 바이트들 중 제 1 바이트 다음에 제 2 바이트가 출현하여, 연결되는 확률일 수 있다.
- [0114] 예를 들어, 바이너리 파일 변환부(120)는 0의 값을 갖는 제 1 바이트 다음에 1의 값을 갖는 제 2 바이트가 출현되는 연결 확률을 계산할 수 있다. 바이너리 파일 변환부(120)는 1의 값을 갖는 제 2 바이트 다음에 2의 값을 갖는 제 3 바이트가 출현되는 연결 확률을 계산할 수 있다.
- [0115] 바이너리 파일 변환부(120)는 연결 확률에 기초하여, 입력 데이터를 기 설정된 크기의 데이터로 변환할 수 있다(S231).
- [0116] 실시예에 따라, 바이너리 파일 변환부(120)는 입력 데이터를 256x256 크기를 갖는 행렬 구조의 데이터로 변환할 수 있다. 다만, 기 설정된 크기는 256x256 크기로 한정되는 것은 아니며, 이보다 더 크거나 또는 더 작은 크기일 수도 있다.
- [0117] 상술한 구성에 따르면, 컴퓨팅 장치(100)는 운영체제 환경에 따라 상이한 바이너리 파일 형태의 입력 데이터를 기 설정된 크기의 데이터로 변환할 수 있다. 여기서, 기 설정된 크기의 데이터는 행렬 구조의 데이터일 수 있다. 이에 따라, 컴퓨팅 장치(100)는 웨이블릿 변환 기법을 이용하여, 기 설정된 크기의 데이터로부터 특징 정보를 추출할 수 있다. 이하에서는 컴퓨팅 장치(100)가 웨이블릿 변환 기법을 이용하여 기 설정된 크기의 데이터로부터 특징 정보를 추출하는 방법에 대해 설명한다.
- [0119] 도 7은 컴퓨팅 장치가 웨이블릿 변환 기법을 이용하여 기 설정된 크기의 데이터로부터 특징 정보를 추출하는 방법의 일례를 설명하기 위한 흐름도이다.
- [0120] 도 7을 참조하면, 컴퓨팅 장치(100)의 신호처리부(130)는 기 설정된 크기의 데이터에 웨이블릿 변환을 수행하여, 웨이블릿 계수 행렬(Wavelet coefficients matrix)을 추출할 수 있다(S310). 웨이블릿 계수는 데이터에 웨이블릿 변환을 수행함에 따라 얻어지는 계수일 수 있다. 웨이블릿 계수 행렬은 웨이블릿 계수를 이용하여 구축되는 행렬일 수 있다.
- [0121] 웨이블릿 계수 행렬은 근사 계수 행렬(approximation coefficients matrix) 및 상세 계수 행렬(detail coefficients matrix)을 포함할 수 있다. 근사 계수 행렬은 기 설정된 크기의 데이터의 저주파수 영역에 대한 정보를 포함할 수 있다. 상세 계수 행렬은 기 설정된 크기의 데이터의 고주파수 영역에 대한 정보를 포함할 수 있다.
- [0122] 본 개시의 몇몇 실시예에 따르면, 상세 계수 행렬은 수평 계수 행렬(horizontal coefficients matrix), 수직 계수 행렬(vertical coefficients matrix) 또는 대각선 계수 행렬(diagonal coefficients matrix) 중 적어도 하나를 포함할 수 있다. 이에 따라, 상세 계수 행렬은 기 설정된 크기의 데이터의 고주파수 영역에 대한 수평 방향의 정보, 기 설정된 크기의 데이터의 고주파수 영역에 대한 수직 방향의 정보 또는 기 설정된 크기의 데이터의 고주파수 영역에 대한 대각선 방향의 정보 중 적어도 하나를 포함할 수 있다.
- [0123] 신호처리부(130)는 웨이블릿 계수 행렬에 기초하여, 입력 데이터의 특징 정보를 추출할 수 있다(S320). 특징 정보는 바이너리 파일 형태의 입력 데이터가 악성코드인지 여부를 탐지하기 위해 사용되는 정보일 수 있다.
- [0124] 특징 정보는 기 설정된 크기의 데이터가 1x1 행렬 구조가 되도록 제 1 횟수의 웨이블릿 변환을 수행하여 추출되는 근사 계수 행렬과 관련된 제 1 특징을 포함할 수 있다. 실시예에 따라, 기 설정된 크기의 데이터는 256x256 크기를 갖는 행렬 구조의 데이터일 수 있다. 따라서, 신호처리부(130)는 8회의 웨이블릿 변환을 수행하여, 기 설정된 크기의 데이터가 1x1 행렬 구조가 되도록 할 수 있다. 다만, 제 1 횟수는 8회에 한정되는 것은 아니며, 다양한 크기를 갖는 행렬 구조의 데이터들이 1x1 행렬 구조가 될 수 있도록 하는 횟수일 수 있다. 일례로, 기 설정된 크기의 데이터가 512x512 크기를 갖는 행렬 구조의 데이터일 경우, 신호처리부(130)는 9회의 웨이블릿 변환을 수행하여, 기 설정된 크기의 데이터가 1x1 행렬 구조가 되도록 할 수 있다.
- [0125] 특징 정보는 제 1 임계 횟수 이상이고 그리고 제 1 횟수 이하의 제 1 개수 범위 내에서의 제 2 횟수의 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬 값에 대한 제 2 특징을 포함할 수 있다. 여기서, 제 1 임계 횟수는 제 1 횟수보다 작은 횟수일 수 있다.

- [0126] 일례로, 제 1 횟수는 8회이고, 제 1 임계 횟수는 4회일 수 있다. 이 경우, 제 1 개수 범위는 8회에서 4회일 수 있고, 제 2 횟수는 8회, 7회, 6회, 5회 및 4회일 수 있다. 따라서, 신호처리부(130)는 기 설정된 크기의 데이터에 8회, 7회, 6회, 5회 및 4회의 웨이블릿 변환을 수행하여 대각선 계수 행렬 값에 대한 제 2 특징을 추출할 수 있다.
- [0127] 다른 일례로, 제 1 횟수는 8회이고, 제 1 임계 횟수는 7회일 수 있다. 이 경우, 제 1 개수 범위는 8회에서 7회일 수 있고, 제 2 횟수는 8회 및 7회일 수 있다. 따라서, 신호처리부(130)는 기 설정된 크기의 데이터에 8회 및 7회의 웨이블릿 변환을 수행하여 대각선 계수 행렬 값에 대한 제 2 특징을 추출할 수 있다.
- [0128] 본 개시의 몇몇 실시예에 따르면 제 1 임계 횟수는 컴퓨팅 장치(100)의 성능에 기초하여 결정될 수 있다. 예를 들어, 신호처리부(130)는 컴퓨팅 장치(100)의 성능이 충분한 경우, 제 1 임계 횟수를 1회로 결정할 수 있다. 제 1 임계 횟수가 1회인 경우, 제 1 개수 범위는 8회에서 1회일 수 있고, 이에 따라 제 2 횟수는 8회, 7회, 6회, 5회, 4회, 3회, 2회 및 1회일 수 있다. 신호처리부(130)가 8회, 7회, 6회, 5회, 4회, 3회, 2회 및 1회의 웨이블릿 변환을 수행하여 대각선 계수 행렬 값을 추출하고, 추출된 대각선 계수 행렬 값에 기초하여 입력 데이터의 특징 정보를 추출하는 경우 특징 정보의 개수가 늘어날 수 있다. 또는 특징 정보가 좀 더 다양해질 수 있다. 따라서, 추출된 특징 정보를 학습 데이터로 이용하여 네트워크 모델을 학습시키는 경우, 학습된 네트워크 모델의 성능이 향상될 수 있다.
- [0129] 다른 예를 들어, 신호처리부(130)는 컴퓨팅 장치(100)의 성능이 충분하지 못한 경우, 제 1 임계 횟수를 4회로 결정할 수 있다. 이 경우, 신호처리부(130)가 기 설정된 크기의 데이터에 웨이블릿 변환을 수행하는 속도가 향상될 수 있고, 또한 리소스가 절약될 수 있다.
- [0130] 특징 정보는 제 1 임계 횟수 이상이고 그리고 제 1 횟수 미만의 제 2 개수 범위 내에서의 제 3 횟수의 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬과 관련된 제 3 특징을 포함할 수 있다.
- [0131] 일례로, 제 1 횟수는 8회이고, 제 1 임계 횟수는 4회일 수 있다. 이 경우, 제 2 개수 범위는 7회에서 4회일 수 있고, 제 3 횟수는 7회, 6회, 5회 및 4회일 수 있다. 따라서, 신호처리부(130)는 기 설정된 크기의 데이터에 7회, 6회, 5회 및 4회의 웨이블릿 변환을 수행하여 대각선 계수 행렬과 관련된 제 3 특징을 추출할 수 있다.
- [0132] 다른 일례로, 제 1 횟수는 8회이고, 제 1 임계 횟수는 6회일 수 있다. 이 경우, 제 2 개수 범위는 7회에서 6회일 수 있고, 제 3 횟수는 7회 및 6회일 수 있다. 따라서, 신호처리부(130)는 기 설정된 크기의 데이터에 7회 및 6회의 웨이블릿 변환을 수행하여 대각선 계수 행렬과 관련된 제 3 특징을 추출할 수 있다.
- [0133] 본 개시의 몇몇 실시예에 따르면, 기 설정된 크기의 데이터에 8회의 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬과 관련된 제 3 특징은 항상 동일한 결과 값이 추출될 수 있다.
- [0134] 구체적으로, 제 3 특징은 제 3 횟수의 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬의 비율과 관련된 제 3-1 특징 또는 제 3 횟수의 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬의 제 1 표준편차와 관련된 제 3-2 특징 중 적어도 하나를 포함할 수 있다. 대각선 계수 행렬의 비율은 기 설정된 크기의 데이터에 웨이블릿 변환을 수행하여 추출되는 웨이블릿 계수 행렬들에 있어서, 대각선 계수 행렬이 차지하는 비율일 수 있다. 대각선 계수 행렬의 제 1 표준편차는 기 설정된 크기의 데이터에 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬에 대한 표준 편차일 수 있다. 그러나, 신호처리부(130)가 기 설정된 크기의 데이터에 8회의 웨이블릿 변환을 수행하여 추출되는 제 3-1 특징과 제 3-2 특징의 결과 값은 항상 고정되어 있을 수 있다. 신호처리부(130)가 기 설정된 크기의 데이터에 8회의 웨이블릿 변환을 수행하여 제 3-1 특징 및 제 3-2 특징을 추출하는 동작은 리소스 면에서 낭비될 수 있다. 따라서, 신호처리부(130)는 제 1 횟수 미만의 제 2 개수 범위 내에서의 제 3 횟수의 웨이블릿 변환을 수행하여 대각선 계수 행렬과 관련된 제 3 특징을 추출할 수 있다.
- [0135] 본 개시의 몇몇 실시예에 따르면 제 1 임계 횟수는 컴퓨팅 장치(100)의 성능에 기초하여 결정될 수 있다. 예를 들어, 신호처리부(130)는 컴퓨팅 장치(100)의 성능이 충분한 경우, 제 1 임계 횟수를 1회로 결정할 수 있다. 제 1 임계 횟수가 1회인 경우, 제 2 개수 범위는 7회에서 1회일 수 있고, 이에 따라 제 3 횟수는 7회, 6회, 5회, 4회, 3회, 2회 및 1회일 수 있다. 다른 예를 들어, 신호처리부(130)는 컴퓨팅 장치(100)의 성능이 충분하지 못한 경우, 제 1 임계 횟수를 4회로 결정할 수 있다. 제 1 임계 횟수가 4회인 경우, 제 2 개수 범위는 7회에서 4회일 수 있고, 이에 따라 제 3 횟수는 7회, 6회, 5회 및 4회일 수 있다.
- [0136] 특징 정보는 제 1 임계 횟수 미만의 제 2 임계 횟수 이상이고 그리고 제 1 횟수 미만의 제 3 개수 범위 내에서의 제 4 횟수의 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬과 관련된 제 4 특징을 포함할 수 있다.

- [0137] 일례로, 제 1 횡수는 8회이고, 제 1 임계 횡수는 7회이고, 제 2 임계 횡수는 1회일 수 있다. 이 경우, 제 2 개수 범위는 7회에서 1회일 수 있고, 제 4 횡수는 7회, 6회, 5회, 4회, 3회, 2회 및 1회일 수 있다. 따라서, 신호처리부(130)는 기 설정된 크기의 데이터에 7회, 6회, 5회, 4회, 3회, 2회 및 1회의 웨이블릿 변환을 수행하여 대각선 계수 행렬과 관련된 제 4 특징을 추출할 수 있다.
- [0138] 본 개시의 몇몇 실시예에 따르면, 기 설정된 크기의 데이터에 8회의 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬과 관련된 제 4 특징은 항상 동일한 결과 값이 추출될 수 있다.
- [0139] 구체적으로, 제 4 특징은 제 4 횡수의 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬의 값의 합 관련된 제 4-1 특징, 제 4 횡수의 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬의 값의 비대칭도와 관련된 제 4-2 특징, 제 4 횡수의 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬의 제 2 표준편차와 관련된 제 4-3 특징, 제 4 횡수의 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬의 총 평균 값과 관련된 제 4-4 특징 또는 제 4 횡수의 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬의 히스토그램 퍼센트와 관련된 제 4-5 특징 중 적어도 하나의 특징을 포함할 수 있다. 대각선 계수 행렬의 값의 합은 기 설정된 크기의 데이터에 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬 값들의 합일 수 있다. 대각선 계수 행렬의 값의 비대칭도는 기 설정된 크기의 데이터에 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬 값들의 비대칭도 일 수 있다. 대각선 계수 행렬의 제 2 표준편차는 기 설정된 크기의 데이터에 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬에 대한 표준 편차일 수 있다. 대각선 계수 행렬의 평균 값은 기 설정된 크기의 데이터에 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬의 전체 값에 대한 평균 값일 수 있다. 그러나, 신호처리부(130)가 기 설정된 크기의 데이터에 8회의 웨이블릿 변환을 수행하여 추출되는 제 4-1 특징과 기 설정된 크기의 데이터에 8회의 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬 값에 대한 제 2 특징은 동일할 수 있다. 신호처리부(130)가 기 설정된 크기의 데이터에 8회의 웨이블릿 변환을 수행하여 추출되는 제 4-2 특징, 제 4-3 특징, 제 4-4 특징 및 제 4-5 특징은 항상 동일한 결과 값이 추출될 수 있다. 구체적으로, 제 4-2 특징은 비대칭도와 관련되고, 제 4-3 특징은 표준편차와 관련되고, 제 4-4 특징은 평균 값과 관련되며, 제 4-5 특징은 히스토그램 퍼센트와 관련되는 특징인데, 웨이블릿 변환을 8회 수행하여 추출되는 결과 값은 항상 한 개의 값만 가지고 있기 때문일 수 있다. 따라서, 신호처리부(130)가 기 설정된 크기의 데이터에 8회의 웨이블릿 변환을 수행하여 제 4 특징을 추출하는 동작은 리소스 면에서 낭비될 수 있다. 그러므로, 신호처리부(130)는 제 4 횡수의 웨이블릿 변환을 수행하여 추출되는 대각선 계수 행렬과 관련된 제 4 특징을 추출할 수 있다.
- [0140] 본 개시에서, 특징 정보는 웨이블릿 변환을 수행하지 않은 기 설정된 크기의 데이터와 관련된 제 5 특징을 포함할 수 있다. 일례로, 제 5 특징은 기 설정된 크기의 데이터의 히스토그램 퍼센트와 관련된 제 5-1 특징, 기 설정된 크기의 데이터의 표준편차와 관련된 제 5-2 특징 및 기 설정된 크기의 데이터의 총 표준편차와 관련된 제 5-3 특징을 포함할 수 있다.
- [0141] 상술한 바와 같이, 신호처리부(130)는 웨이블릿 계수 행렬에 기초하여, 입력 데이터로부터 다양한 종류의 특징 정보를 추출할 수 있다. 추출된 특징 정보는 입력 데이터로부터 악성 코드를 탐지하도록 훈련되는 네트워크 모델에 학습 데이터로 이용될 수 있다. 따라서, 네트워크 모델은 높은 정확도로 악성 코드를 탐지 및 예측할 수 있도록 훈련될 수 있다.
- [0143] 도 8은 본 개시내용의 실시예들이 구현될 수 있는 예시적인 컴퓨팅 환경에 대한 일반적인 개략도를 도시한다.
- [0144] 본 개시내용이 일반적으로 하나 이상의 컴퓨터 상에서 실행될 수 있는 컴퓨터 실행가능 명령어와 관련하여 기술되었지만, 당업자라면 본 개시내용 기타 프로그램 모듈들과 결합되어 및/또는 하드웨어와 소프트웨어의 조합으로서 구현될 수 있다는 것을 잘 알 것이다.
- [0145] 일반적으로, 본 명세서에서의 모듈은 특정의 태스크를 수행하거나 특정의 추상 데이터 유형을 구현하는 루틴, 프로시저, 프로그램, 컴포넌트, 데이터 구조, 기타 등등을 포함한다. 또한, 당업자라면 본 개시의 방법이 단일-프로세서 또는 멀티프로세서 컴퓨터 시스템, 미니컴퓨터, 메인프레임 컴퓨터는 물론 퍼스널 컴퓨터, 핸드헬드 컴퓨팅 장치, 마이크로프로세서-기반 또는 프로그램가능 가전 제품, 기타 등등(이들 각각은 하나 이상의 연관된 장치와 연결되어 동작할 수 있음)을 비롯한 다른 컴퓨터 시스템 구성으로 실시될 수 있다는 것을 잘 알 것이다.
- [0146] 본 개시의 설명된 실시예들은 또한 어떤 태스크들이 통신 네트워크를 통해 연결되어 있는 원격 처리 장치들에 의해 수행되는 분산 컴퓨팅 환경에서 실시될 수 있다. 분산 컴퓨팅 환경에서, 프로그램 모듈은 로컬 및 원격 메모리 저장 장치 둘다에 위치할 수 있다.

- [0147] 컴퓨터는 통상적으로 다양한 컴퓨터 판독가능 매체를 포함한다. 컴퓨터에 의해 액세스 가능한 매체로서, 휘발성 및 비휘발성 매체, 일시적(transitory) 및 비일시적(non-transitory) 매체, 이동식 및 비-이동식 매체를 포함한다. 제한이 아닌 예로서, 컴퓨터 판독가능 매체는 컴퓨터 판독가능 저장 매체 및 컴퓨터 판독가능 전송 매체를 포함할 수 있다.
- [0148] 컴퓨터 판독가능 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보를 저장하는 임의의 방법 또는 기술로 구현되는 휘발성 및 비휘발성 매체, 일시적 및 비-일시적 매체, 이동식 및 비이동식 매체를 포함한다. 컴퓨터 판독가능 저장 매체는 RAM, ROM, EEPROM, 플래시 메모리 또는 기타 메모리 기술, CD-ROM, DVD(digital video disk) 또는 기타 광 디스크 저장 장치, 자기 카세트, 자기 테이프, 자기 디스크 저장 장치 또는 기타 자기 저장 장치, 또는 컴퓨터에 의해 액세스될 수 있고 원하는 정보를 저장하는데 사용될 수 있는 임의의 기타 매체를 포함하지만, 이에 한정되지 않는다.
- [0149] 컴퓨터 판독가능 전송 매체는 통상적으로 반송파(carrier wave) 또는 기타 전송 메커니즘(transport mechanism)과 같은 피변조 데이터 신호(modulated data signal)에 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터들을 구현하고 모든 정보 전달 매체를 포함한다. 피변조 데이터 신호라는 용어는 신호 내에 정보를 인코딩하도록 그 신호의 특성들 중 하나 이상을 설정 또는 변경시킨 신호를 의미한다. 제한이 아닌 예로서, 컴퓨터 판독가능 전송 매체는 유선 네트워크 또는 직접 배선 접속(direct-wired connection)과 같은 유선 매체, 그리고 음향, RF, 적외선, 기타 무선 매체와 같은 무선 매체를 포함한다. 상술된 매체들 중 임의의 것의 조합도 역시 컴퓨터 판독가능 전송 매체의 범위 안에 포함되는 것으로 한다.
- [0150] 컴퓨터(1102)를 포함하는 본 개시의 여러가지 측면들을 구현하는 예시적인 환경(1100)이 나타내어져 있으며, 컴퓨터(1102)는 처리 장치(1104), 시스템 메모리(1106) 및 시스템 버스(1108)를 포함한다. 시스템 버스(1108)는 시스템 메모리(1106)(이에 한정되지 않음)를 비롯한 시스템 컴포넌트들을 처리 장치(1104)에 연결시킨다. 처리 장치(1104)는 다양한 상용 프로세서들 중 임의의 프로세서일 수 있다. 듀얼 프로세서 및 기타 멀티프로세서 아키텍처도 역시 처리 장치(1104)로서 이용될 수 있다.
- [0151] 시스템 버스(1108)는 메모리 버스, 주변장치 버스, 및 다양한 상용 버스 아키텍처 중 임의의 것을 사용하는 로컬 버스에 추가적으로 상호 연결될 수 있는 몇 가지 유형의 버스 구조 중 임의의 것일 수 있다. 시스템 메모리(1106)는 판독 전용 메모리(ROM)(1110) 및 랜덤 액세스 메모리(RAM)(1112)를 포함한다. 기본 입/출력 시스템(BIOS)은 ROM, EPROM, EEPROM 등의 비휘발성 메모리(1110)에 저장되며, 이 BIOS는 시동 중과 같은 때에 컴퓨터(1102) 내의 구성요소들 간에 정보를 전송하는 일을 돕는 기본적인 루틴을 포함한다. RAM(1112)은 또한 데이터를 캐싱하기 위한 정적 RAM 등의 고속 RAM을 포함할 수 있다.
- [0152] 컴퓨터(1102)는 또한 내장형 하드 디스크 드라이브(HDD)(1114)(예를 들어, EIDE, SATA)–이 내장형 하드 디스크 드라이브(1114)는 또한 적당한 새시(도시 생략) 내에서 외장형 용도로 구성될 수 있음–, 자기 플로피 디스크 드라이브(FDD)(1116)(예를 들어, 이동식 디스켓(1118)으로부터 판독을 하거나 그에 기록을 하기 위한 것임), 및 광 디스크 드라이브(1120)(예를 들어, CD-ROM 디스크(1122)를 판독하거나 DVD 등의 기타 고용량 광 매체로부터 판독을 하거나 그에 기록을 하기 위한 것임)를 포함한다. 하드 디스크 드라이브(1114), 자기 디스크 드라이브(1116) 및 광 디스크 드라이브(1120)는 각각 하드 디스크 드라이브 인터페이스(1124), 자기 디스크 드라이브 인터페이스(1126) 및 광 드라이브 인터페이스(1128)에 의해 시스템 버스(1108)에 연결될 수 있다. 외장형 드라이브 구현을 위한 인터페이스(1124)는 예를 들어, USB(Universal Serial Bus) 및 IEEE 1394 인터페이스 기술 중 적어도 하나 또는 그 둘 다를 포함한다.
- [0153] 이들 드라이브 및 그와 연관된 컴퓨터 판독가능 매체는 데이터, 데이터 구조, 컴퓨터 실행가능 명령어, 기타 등등의 비휘발성 저장을 제공한다. 컴퓨터(1102)의 경우, 드라이브 및 매체는 임의의 데이터를 적당한 디지털 형식으로 저장하는 것에 대응한다. 상기에서의 컴퓨터 판독가능 저장 매체에 대한 설명이 HDD, 이동식 자기 디스크, 및 CD 또는 DVD 등의 이동식 광 매체를 언급하고 있지만, 당업자라면 zip 드라이브(zip drive), 자기 카세트, 플래시 메모리 카드, 카트리지, 기타 등등의 컴퓨터에 의해 판독가능한 다른 유형의 저장 매체도 역시 예시적인 운영 환경에서 사용될 수 있으며 또 임의의 이러한 매체가 본 개시의 방법들을 수행하기 위한 컴퓨터 실행가능 명령어를 포함할 수 있다는 것을 잘 알 것이다.
- [0154] 운영 체제(1130), 하나 이상의 애플리케이션 프로그램(1132), 기타 프로그램 모듈(1134) 및 프로그램 데이터(1136)를 비롯한 다수의 프로그램 모듈이 드라이브 및 RAM(1112)에 저장될 수 있다. 운영 체제, 애플리케이션, 모듈 및/또는 데이터의 전부 또는 그 일부분이 또한 RAM(1112)에 캐싱될 수 있다. 본 개시가 여러가지 상업적으

로 이용가능한 운영 체제 또는 운영 체제들의 조합에서 구현될 수 있다는 것을 잘 알 것이다.

- [0155] 사용자는 하나 이상의 유선/무선 입력 장치, 예를 들어, 키보드(1138) 및 마우스(1140) 등의 포인팅 장치를 통해 컴퓨터(1102)에 명령 및 정보를 입력할 수 있다. 기타 입력 장치(도시 생략)로는 마이크, IR 리모콘, 조이스틱, 게임 패드, 스타일러스 펜, 터치 스크린, 기타 등등이 있을 수 있다. 이들 및 기타 입력 장치가 종종 시스템 버스(1108)에 연결되어 있는 입력 장치 인터페이스(1142)를 통해 처리 장치(1104)에 연결되지만, 병렬 포트, IEEE 1394 직렬 포트, 게임 포트, USB 포트, IR 인터페이스, 기타 등등의 기타 인터페이스에 의해 연결될 수 있다.
- [0156] 모니터(1144) 또는 다른 유형의 디스플레이 장치도 역시 비디오 어댑터(1146) 등의 인터페이스를 통해 시스템 버스(1108)에 연결된다. 모니터(1144)에 부가하여, 컴퓨터는 일반적으로 스피커, 프린터, 기타 등등의 기타 주변 출력 장치(도시 생략)를 포함한다.
- [0157] 컴퓨터(1102)는 유선 및/또는 무선 통신을 통한 원격 컴퓨터(들)(1148) 등의 하나 이상의 원격 컴퓨터로의 논리적 연결을 사용하여 네트워크화된 환경에서 동작할 수 있다. 원격 컴퓨터(들)(1148)는 워크스테이션, 서버 컴퓨터, 라우터, 퍼스널 컴퓨터, 휴대용 컴퓨터, 마이크로프로세서-기반 오락 기기, 피어 장치 또는 기타 통상의 네트워크 노드일 수 있으며, 일반적으로 컴퓨터(1102)에 대해 기술된 구성요소들 중 다수 또는 그 전부를 포함하지만, 간략함을 위해, 메모리 저장 장치(1150)만이 도시되어 있다. 도시되어 있는 논리적 연결은 근거리 통신망(LAN)(1152) 및/또는 더 큰 네트워크, 예를 들어, 원격 통신망(WAN)(1154)에의 유선/무선 연결을 포함한다. 이러한 LAN 및 WAN 네트워킹 환경은 사무실 및 회사에서 일반적인 것이며, 인트라넷 등의 전사적 컴퓨터 네트워크(enterprise-wide computer network)를 용이하게 해주며, 이들 모두는 전세계 컴퓨터 네트워크, 예를 들어, 인터넷에 연결될 수 있다.
- [0158] LAN 네트워킹 환경에서 사용될 때, 컴퓨터(1102)는 유선 및/또는 무선 통신 네트워크 인터페이스 또는 어댑터(1156)를 통해 로컬 네트워크(1152)에 연결된다. 어댑터(1156)는 LAN(1152)에의 유선 또는 무선 통신을 용이하게 해줄 수 있으며, 이 LAN(1152)은 또한 무선 어댑터(1156)와 통신하기 위해 그에 설치되어 있는 무선 액세스 포인트를 포함하고 있다. WAN 네트워킹 환경에서 사용될 때, 컴퓨터(1102)는 모뎀(1158)을 포함할 수 있거나, WAN(1154) 상의 통신 서버에 연결되거나, 또는 인터넷을 통하는 등, WAN(1154)을 통해 통신을 설정하는 기타 수단을 갖는다. 내장형 또는 외장형 및 유선 또는 무선 장치일 수 있는 모뎀(1158)은 직렬 포트 인터페이스(1142)를 통해 시스템 버스(1108)에 연결된다. 네트워크화된 환경에서, 컴퓨터(1102)에 대해 설명된 프로그램 모듈들 또는 그의 일부분이 원격 메모리/저장 장치(1150)에 저장될 수 있다. 도시된 네트워크 연결이 예시적인 것이며 컴퓨터들 사이에 통신 링크를 설정하는 기타 수단이 사용될 수 있다는 것을 잘 알 것이다.
- [0159] 컴퓨터(1102)는 무선 통신으로 배치되어 동작하는 임의의 무선 장치 또는 개체, 예를 들어, 프린터, 스캐너, 데스크톱 및/또는 휴대용 컴퓨터, PDA(portable data assistant), 통신 위성, 무선 검출가능 태그와 연관된 임의의 장비 또는 장소, 및 전화와 통신을 하는 동작을 한다. 이것은 적어도 Wi-Fi 및 블루투스 무선 기술을 포함한다. 따라서, 통신은 종래의 네트워크에서와 같이 미리 정의된 구조이거나 단순히 적어도 2개의 장치 사이의 애드혹 통신(ad hoc communication)일 수 있다.
- [0160] Wi-Fi(Wireless Fidelity)는 유선 없이도 인터넷 등으로의 연결을 가능하게 해준다. Wi-Fi는 이러한 장치, 예를 들어, 컴퓨터가 실내에서 및 실외에서, 즉 기지국의 통화권 내의 아무 곳에서나 데이터를 전송 및 수신할 수 있게 해주는 셀 전화와 같은 무선 기술이다. Wi-Fi 네트워크는 안전하고 신뢰성 있으며 고속인 무선 연결을 제공하기 위해 IEEE 802.11(a,b,g, 기타)이라고 하는 무선 기술을 사용한다. 컴퓨터를 서로에, 인터넷에 및 유선 네트워크(IEEE 802.3 또는 이더넷을 사용함)에 연결시키기 위해 Wi-Fi가 사용될 수 있다. Wi-Fi 네트워크는 비인가 2.4 및 5 GHz 무선 대역에서, 예를 들어, 11Mbps(802.11a) 또는 54 Mbps(802.11b) 데이터 레이트로 동작하거나, 양 대역(듀얼 대역)을 포함하는 제품에서 동작할 수 있다.
- [0161] 본 개시의 기술 분야에서 통상의 지식을 가진 자는 여기에 개시된 실시예들과 관련하여 설명된 다양한 예시적인 논리 블록들, 모듈들, 프로세서들, 수단들, 회로들 및 알고리즘 단계들이 전자 하드웨어, (편의를 위해, 여기에서 "소프트웨어"로 지칭되는) 다양한 형태들의 프로그램 또는 설계 코드 또는 이들 모두의 결합에 의해 구현될 수 있다는 것을 이해할 것이다. 하드웨어 및 소프트웨어의 이러한 상호 호환성을 명확하게 설명하기 위해, 다양한 예시적인 컴포넌트들, 블록들, 모듈들, 회로들 및 단계들이 이들의 기능과 관련하여 위에서 일반적으로 설명되었다. 이러한 기능이 하드웨어 또는 소프트웨어로서 구현되는지 여부는 특정한 애플리케이션 및 전체 시스템에 대하여 부과되는 설계 제약들에 따라 좌우된다. 본 개시의 기술 분야에서 통상의 지식을 가진 자는 각각의 특정한 애플리케이션에 대하여 다양한 방식으로 설명된 기능을 구현할 수 있으나, 이러한 구현 결정들은 본 개

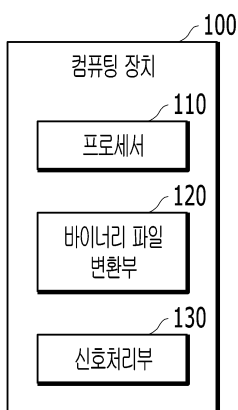
시의 범위를 벗어나는 것으로 해석되어서는 안 될 것이다.

[0162] 여기서 제시된 다양한 실시예들은 방법, 장치, 또는 표준 프로그래밍 및/또는 엔지니어링 기술을 사용한 제조 물품(article)으로 구현될 수 있다. 용어 "제조 물품"은 임의의 컴퓨터-판독가능 장치로부터 액세스 가능한 컴퓨터 프로그램 또는 매체(media)를 포함한다. 예를 들어, 컴퓨터-판독가능 저장 매체는 자기 저장 장치(예를 들면, 하드 디스크, 플로피 디스크, 자기 스트립, 등), 광학 디스크(예를 들면, CD, DVD, 등), 스마트 카드, 및 플래쉬 메모리 장치(예를 들면, EEPROM, 카드, 스틱, 키 드라이브, 등)를 포함하지만, 이들로 제한되는 것은 아니다. 용어 "기계-판독가능 매체"는 명령(들) 및/또는 데이터를 저장, 보유, 및/또는 전달할 수 있는 무선 채널 및 다양한 다른 매체를 포함하지만, 이들로 제한되는 것은 아니다.

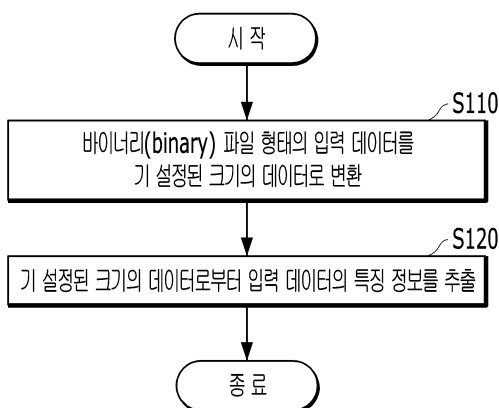
[0164] 제시된 실시예들에 대한 설명은 임의의 본 개시의 기술 분야에서 통상의 지식을 가진 자가 본 개시를 이용하거나 또는 실시할 수 있도록 제공된다. 이러한 실시예들에 대한 다양한 변형들은 본 개시의 기술 분야에서 통상의 지식을 가진 자에게 명백할 것이며, 여기에 정의된 일반적인 원리들은 본 개시의 범위를 벗어남이 없이 다른 실시예들에 적용될 수 있다. 그리하여, 본 개시는 여기에 제시된 실시예들로 한정되는 것이 아니라, 여기에 제시된 원리들 및 신규한 특징들과 일관되는 최광의 범위에서 해석되어야 할 것이다.

도면

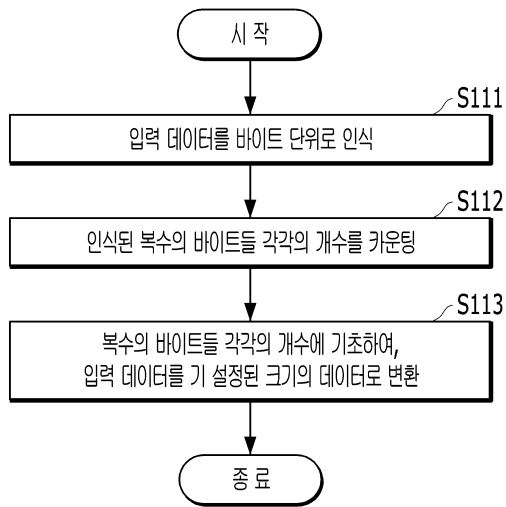
도면1



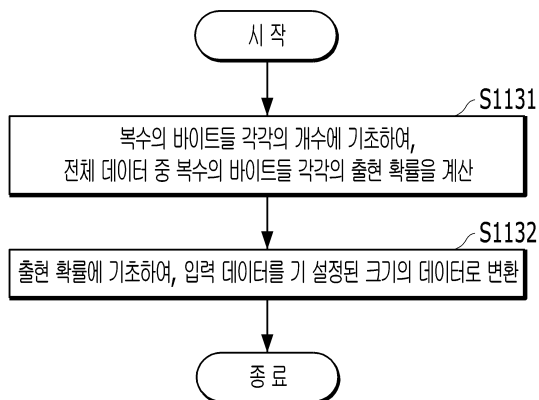
도면2



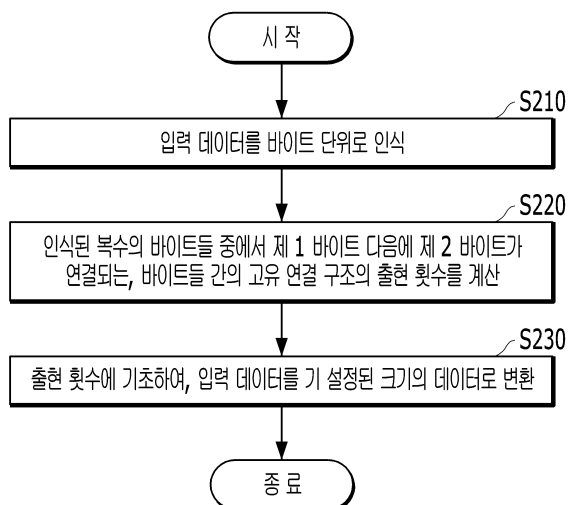
도면3



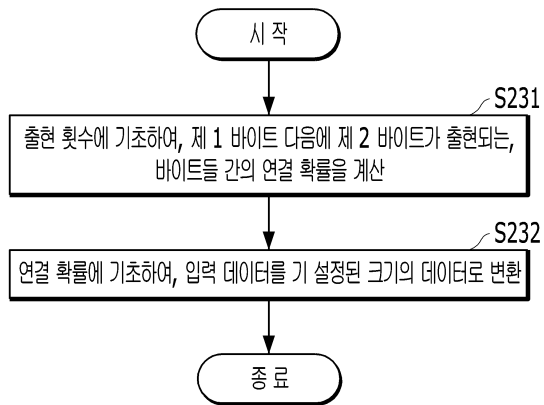
도면4



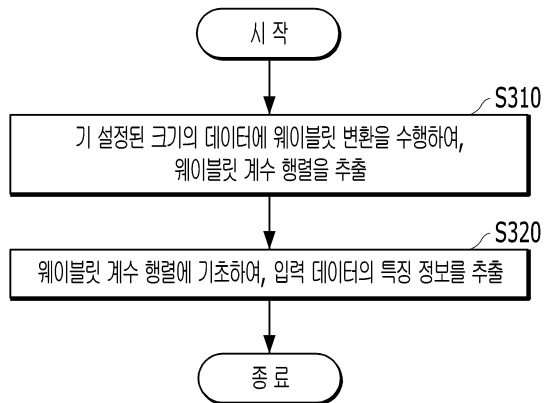
도면5



도면6



도면7



도면8

